CK-12 Foundation

Anoka Hennepin Probability and Statistics

Say Thanks to the Authors
Click http://www.ck12.org/saythanks
(No sign in required)

Engelhaupt **Haney** **Hennepin** **Johnson**

To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Copyright © 2011 CK-12 Foundation, www.ck12.org

The names "CK-12" and "CK12" and associated logos and the terms "**FlexBook®**," and "**FlexBook Platform®**," (collectively "CK-12 Marks") are trademarks and service marks of CK-12 Foundation and are protected by federal, state, and international laws.

Any form of reproduction of this book in any format or medium, in whole or in sections must include the referral attribution link **http://www.ck12.org/saythanks** (placed in a visible location) in addition to the following terms.

Except as otherwise noted, all CK-12 Content (including CK-12 Curriculum Material) is made available to Users in accordance with the Creative Commons Attribution/Non-Commercial/Share Alike 3.0 Unported (CC-BY-NC-SA) License (http://creativecommons.org/licenses/by-nc-sa/3.0/), as amended and updated by Creative Commons from time to time (the "CC License"), which is incorporated herein by this reference.

Complete terms can be found at http://www.ck12.org/terms.

Printed: July 2, 2012

flexbook
next generation textbooks

CC BY NC SA

Authors

Mr. Michael Engelhaupt, Ms. Heather Haney, Anoka Hennepin, Mr. Ernie Johnson

Contributors

Mr. Bruce DeWitt, Ms. Anne Roehrich, Mr. Tom Skoglund

Editors

Ms. Katie Bruck, Ms. Wendy Durant, Mr. Charles Nowariak, Ms. Julie Rydberg, Ms. Meghann Witchger

Contents

Foreword iv

Preface v

1 Counting Methods 1

1.1 Sample Spaces, Events, and Outcomes . . . 1
1.2 Fundamental Counting Principle . . . 5
1.3 Permutations . . . 11
1.4 Combinations . . . 15
1.5 Mixed Combinations and Permutations . . . 19
1.6 Chapter 1 Review . . . 23

2 Calculating Probabilities 26

2.1 Calculating Basic Probabilities . . . 26
2.2 Compound and Independent Events . . . 37
2.3 Mutually Exclusive Outcomes . . . 45
2.4 Tree Diagrams and Probability Models . . . 52
2.5 Conditional Probabilities and 2-Way Tables . . . 59
2.6 Chapter 2 Review . . . 67

3 Expected Values & Simulation 75

3.1 Probability Models & Expected Value . . . 75
3.2 Applied Expected Value Calculations . . . 85
3.3 Simulation and Experimental Probability . . . 92
3.4 Chapter 3 Review . . . 99

4 Data Collection 102

4.1 DATA . . . 102
4.2 Sample Survey and Census . . . 112
4.3 Random Selection . . . 128

4.4 Statistical Conclusions . . . 134
4.5 Experiments and Observational Studies . . . 140
4.6 Chapter 4 Review . . . 148

5 Analyzing Univariate Data **154**
5.1 Categorical Data . . . 154
5.2 Time Plots & Measures of Central Tendency . . . 167
5.3 Numerical Data: Dot Plots & Stem Plots . . . 178
5.4 Numerical Data: Histograms . . . 190
5.5 Numerical Data: Box Plots & Outliers . . . 198
5.6 Numerical Data: Comparing Data Sets . . . 214
5.7 Chapter 5 Review . . . 225

6 Analyzing Bivariate Data **239**
6.1 Displaying Bivariate Data . . . 239
6.2 Correlation . . . 253
6.3 Least-Squares Regression . . . 264
6.4 More Least-Squares Regression . . . 275
6.5 Chapter 6 Review . . . 281

7 The Normal Distribution **285**
7.1 Introduction to the Normal Curve . . . 285
7.2 Z-Scores, Percentiles, and Normal CDF . . . 295
7.3 Inverse Normal Calculations . . . 303
7.4 Chapter 7 Review . . . 309

8 Appendices **314**
8.1 Appendix A - Tables . . . 314
8.2 Appendix B - Glossary and Index . . . 320
8.3 Appendix C - Calculator Help . . . 333

Foreword

Anoka-Hennepin Schools is fortunate to have many experienced math teachers contribute to this project. These primary authors worked tirelessly during the summer to complete the formidable task of completing a textbook in 60 days.

Meet the Authors

Michael Engelhaupt has taught high school mathematics for 9 years. He received a Master's of Math Education from the University of Minnesota. Currently, Michael teaches Statistics and AP Statistics at Blaine High School in Blaine, MN.

Heather Haney has taught high school mathematics for 19 years and currently teaches at Coon Rapids High School in Coon Rapid, MN. Heather teaches Probability and Algebra.

Ernie Johnson currently teach mathematics at Andover High School in Andover, MN. He has taught mathematics for 19 years teaching courses varying from Algebra, to Statistics, to AP Calculus. Ernie graduated from the University of Minnesota in 1992 with a B.S. in Mathematics and received an M.Ed. in Instructional Systems and Technology in 1998 from the University of Minnesota.

Preface

About the Book

Anoka-Hennepin Schools is thrilled to release the publication of its very own Probability and Statistics textbook. *Anoka-Hennepin Probability and Statistics* (First Edition) represents the work of a large team of dedicated writers and editors who have produced a truly unique and flexible "ebook." Available in multiple electronic formats, the content demonstrates 21st century math learning at is finest. Students can access the book from a CD-ROM, DVD, flash drive, or mobile device like the Kindle or ipod. Access is also available through the web anywhere and anytime in multiple formats.

Technology

While paper copies are available for classroom use, the ebook is interactive and includes web site links, simulations, and real world statistical examples. Students can access the textbook through the district Learning Management Site Moodle where large amounts of supplemental and enrichment content can also be found.

The ebook incorporates the use of the TI 83/84 graphing calculators and students work with spreadsheet software to display and manipulate statistical data. Additional content is available through Kahn Academy, which offers individualized problem activities with instructional videos. Find the ebook @ Http://moodle.anoka.k12.mn.us.

Coverage

This foundational course covers the Minnesota Data, Analysis, and Probability benchmarks. The course also meets Anoka-Hennepin math graduation requirements.

Goals

From the Twins to the weather forecast statistics are used everywhere in our lives. *Anoka-Hennepin Probability and Statistics* demonstrates the connection between statistics and our real world.

Students- Read and immerse yourself in this interactive textbook. Challenge yourself to dig deeper into the content or find solutions to your questions online. This textbook is alive and responsive to your needs. Give feedback to your teacher for incorporation into later revisions. Your input is valued going forward. Enjoy.

Chapter 1

Counting Methods

1.1 Sample Spaces, Events, and Outcomes



Learning Objectives

- Determine the sample space for a given event or series of events
- Produce an organized list of outcomes within a sample space

A **sample space** is a list of all the possible outcomes that may occur. What might happen when you flip a coin? You will either get heads or tails. What will happen when you roll a single die? You will either get a 1, 2, 3, 4, 5, or 6. The sample space for flipping a coin is S={heads, tails}. The sample space for rolling a die is S={1,2,3,4,5,6}

On a coin flip, there are two **outcomes**, heads and tails. There are six different outcomes when considering the **event** of rolling a single die.

Example 1

Suppose you roll two dice. Build a 6 by 6 grid to show the different outcomes that might happen when you add the two dice together.

a) What is the sample space for the different sums that you might get?

b) What is the event for this situation?

c) Based on your grid, which outcome occurs most often?

Solution

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

a) The sample space is S={2,3,4,5,6,7,8,9,10,11,12}.

b) The event is the rolling of the two dice.

c) Notice that a total of 7 can occur 6 different ways. A total of 7 is the most likely outcome.

Example 2

A child orders breakfast at a restaurant. The restaurant has two choices of drinks: milk and orange juice. The restaurant also has three choices of meat: sausage, ham, and bacon. Suppose the child orders one drink and one type of meat.

a) Give the sample space that shows all the different outcomes for what the child might order.

b) How many different outcomes are possible?



Solution

a) For the drinks, use M=Milk and O=Orange Juice. For the meat, use S=Sausage, H=Ham, and B=Bacon. The child might order MS, MH, MB, OS, OH, or OB. The sample space is S={MS, MH, MB, OS, OH, OB}. This list can also be generated using a simple grid as shown below.

	Milk	Orange Juice
Sausage	MS	OS
Ham	MH	OH
Bacon	MB	OB

b) There are six possible outcomes. This can be found simply by counting the number of results within the sample space.

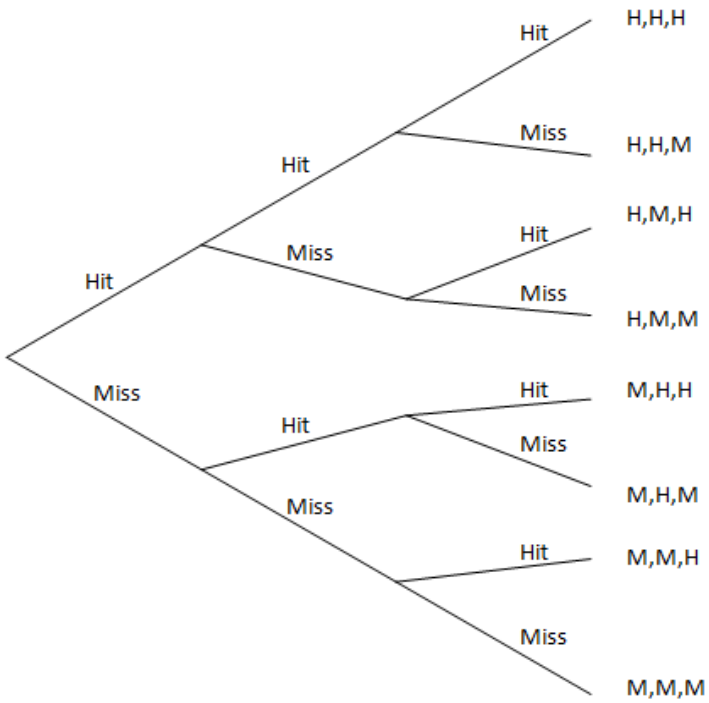
Sometimes, situations can get a bit too complex to simply make a list or build a grid. A **tree diagram** is a visual organizer that is very effective in handling situations with larger numbers of outcomes. We will introduce this concept here, but we will revisit tree diagrams in greater detail in section 1.2.

Example 3

A dart player is trying to hit the bulls-eye with each of three darts that he will throw. Each dart will either hit the bulls-eye or miss the bulls-eye. Use a tree diagram to give the sample space for the different outcomes that may occur.

Solution

Build the tree diagram shown below to track what might happen.



The sample space is S={HHH, HHM, HMH, HMM, MHH, MHM, MMH, MMM}

Problem Set 1.1

1) A single coin is flipped two times.

a) Construct the sample space for this situation.

b) How many different outcomes are possible?

2) A single coin is flipped three times.

a) Use a tree diagram to construct the sample space for this situation.

b) How many different outcomes are possible?

3) A single coin is flipped four times.

a) Construct the sample space for this situation.

b) How many different outcomes are possible?

4) Suppose a 4-sided die is rolled one time. What is the sample space for the result of the roll?

5) Suppose two 4-sided dice are rolled and we keep track of the total on the two dice.

a) Draw a four by four grid that demonstrates the different results for the total of the two dice.

b) What is the sample space for the possible totals of the two dice?



6) Suppose a 4-sided die is rolled two times and we keep track of the *product* when the result from the first die is multiplied by the result from the second die.

a) Draw a four by four grid that demonstrates the different results for the product of the two dice.

b) What is the sample space for the possible products of the two dice?

c) How many different outcomes are possible for the product of the two dice?

d) What outcome occurs most often?

1.2 Fundamental Counting Principle



Learning Objectives

- Apply the Fundamental Counting Principle to determine the number of outcomes
- Create tree diagrams to represent outcomes for a series of events

The **Fundamental Counting Principle** states that if you wish to find the number of outcomes for a given situation, simply multiply the number of outcomes for each individual event. In Example 2 in section 1.1, the child had two different choices of drink and three different choices of meat. If we multiply 2 times 3, we get 6 which is the total number of outcomes possible. The Fundamental Counting Principle expands to any number of events. For example, suppose it turned out that the child also wanted to order eggs and had a choice between scrambled and sunny-side up. The fundamental counting principle states that there are $2 \times 3 \times 2$ or 12 ways to order this breakfast.

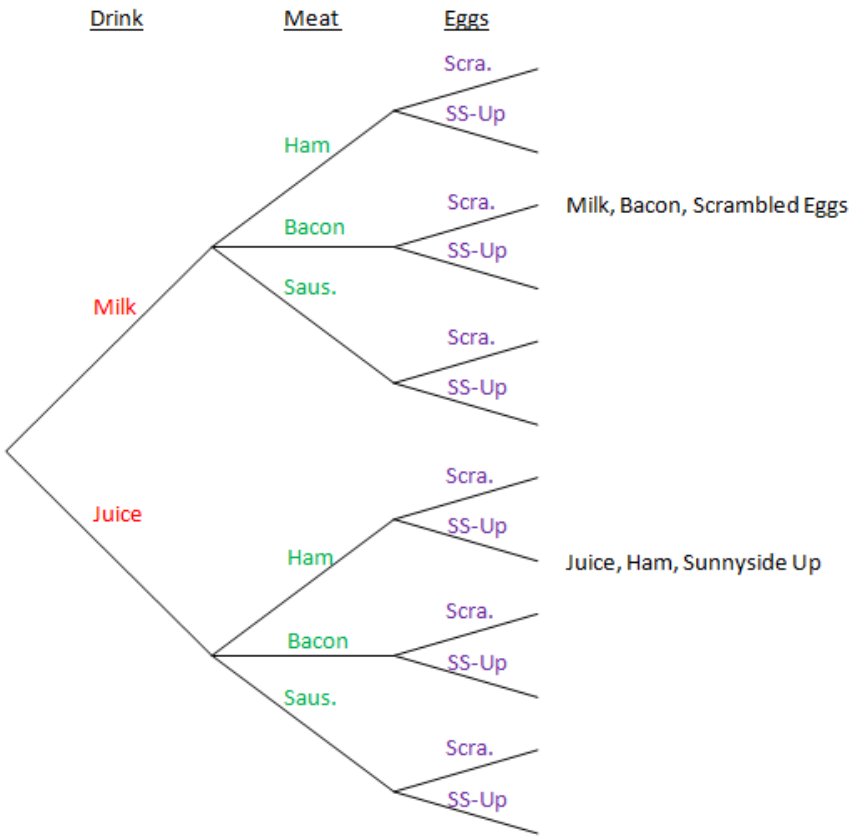
There are other ways to visually see what is happening here. Let's use a **tree diagram**.

Example 1

Build a tree diagram that shows the different outcomes for what the child might order for breakfast.

Solution

The first set of branches of the tree diagram will represent the type of drink, the second set of branches will represent the type of meat, and the third set of branches will represent the type of egg.



We have labeled the ends of two of the branches in the figure on the previous page to show what each branch means. For example, one of the labeled branches shows that the child might have ordered milk, bacon, and scrambled eggs.

The Fundamental Counting Principle is critically important especially when considering complex tree diagrams. Our tree diagram above has many branches and it tracks a great deal of material. It ultimately shows us the 12 different possible breakfast orders, but it takes a large amount of organization to successfully complete. Multiplying 2 by 3 by 2 is a much quicker way to find out the total number of possible outcomes.

Example 2

A couple is planning to have 3 children. Consider the different results that might occur in terms of gender. For example one outcome might be Boy, Boy, Girl (BBG).

a) Using the Fundamental Counting Principle, calculate the number of different outcomes for the children in this family.

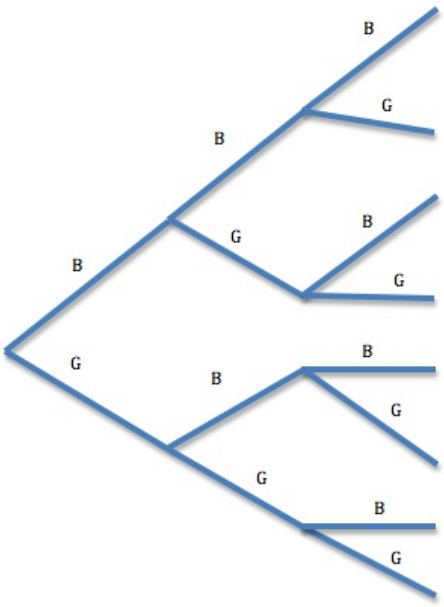
b) Build a tree diagram that shows the different orders of children the couple might have.

c) Construct the sample space that shows all the different orders of children the couple might have.

Solution

a) There are 2 choices for the first child, 2 for the second, and 2 for the third. Therefore, there are $2 \times 2 \times 2 = 8$ outcomes for the gender order of the 3 children.

b)



c) In order to be organized, the list will be alphabetized. BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG There are a total of 8 outcomes.

There are many other ways to apply the Fundamental Counting Principle. A standard deck of cards has 52 cards as shown below. If you are dealt just one card, there are 52 different outcomes.

Standard Deck of 52 Playing Cards

Clubs	Spades	Hearts	Diamonds
A♣	A♠	A♥	A♦
2♣	2♠	2♥	2♦
3♣	3♠	3♥	3♦
4♣	4♠	4♥	4♦
5♣	5♠	5♥	5♦
6♣	6♠	6♥	6♦
7♣	7♠	7♥	7♦
8♣	8♠	8♥	8♦
9♣	9♠	9♥	9♦
10♣	10♠	10♥	10♦
Jack♣	Jack♠	Jack♥	Jack♦
Queen♣	Queen♠	Queen♥	Queen♦
King♣	King♠	King♥	King♦

Example 3

Suppose you are dealt two cards from a standard deck of 52 cards. How many different outcomes are possible?

Solution

We could certainly try drawing a tree diagram but that could get very large quite quickly. The first split alone would have 52 branches on it. On the other hand, if we use the Fundamental Counting Principle, we can simply calculate how many different ways we could be dealt 2 cards from a standard deck. There would be 52 choices for the 1st card and 51 choices for the 2nd card. (Once the first card is dealt, the deck only has 51 cards left in it.) There are $52 \times 51 = 2652$ ways that we could be dealt two cards from the deck.

Example 4

How many different 7-digit phone numbers are possible if no phone number may begin with a zero?

Solution

There are a total of 10 digits available {0, 1, 2, ... 7, 8, 9}. We can't use zero for the first digit so there are only 9 choices for the 1st digit. After that, there are 10 digits available for each of the remaining six digits. This gives us $9 \times 10 \times 10 \times 10 \times 10 \times 10 \times 10 = 9 \times 10^6 = 9,000,000$ ways to come up with a 7 digit phone number.

Example 5

A teenager is given 5 different jobs that they must do before they may go out to a movie with friends. The jobs are washing the car, starting a load of laundry, vacuuming the family room, taking out the garbage, and putting away the dishes. In how many different orders could the teenager complete these jobs?

Solution

There are five choices the teenager could pick for the first job. Once that job is finished, there are only 4 jobs remaining. Once the 2nd job is completed, there are only 3 choices for the 3rd job. Once the 3rd job is finished, there are only 2 choices for the 4th job and finally there will only be one choice left for the 5th job. There are $5 \times 4 \times 3 \times 2 \times 1 = 120$ different orders that these jobs could be completed. Note that there is a quick way to do this ordered multiplication using **factorials**. $5 \times 4 \times 3 \times 2 \times 1 = 5!$ The $5!$ is read "Five Factorial". Be sure to locate the factorial key on your calculator.

Problem Set 1.2

Exercises

1) A woman has three skirts, five shirts, and four hats. How many different outfits can she wear if she picks one skirt, one shirt, and one hat for her outfit?

2) How many different five-digit ZIP codes are possible if the digits can be repeated?

3) How many different five-digit ZIP codes are possible if the digits cannot be repeated?

4) In how many ways can a baseball manager arrange a batting order of nine players?

5) A store manager wishes to display six different kinds of laundry soap by lining them up in a row on a shelf. In how many ways can this be done?

6) There are eight different statistics books, six different geometry books, and three different trigonometry books. In how many ways can a textbook committee select one of each book for their school?

7) At a film festival, there are eight different films that will be shown. In how many different orders can these films be shown?

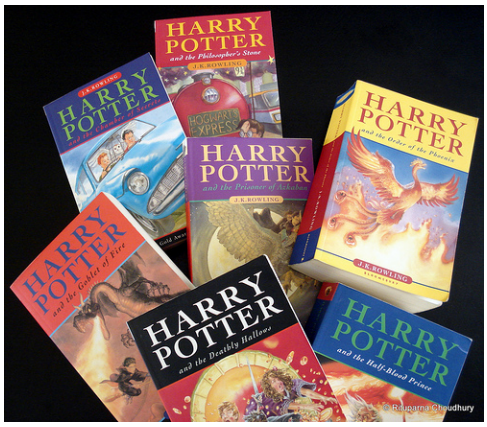
8) The call letters of a radio station must have four letters. The first letter must be a K or a W. How many different call letter combinations are possible if letters may not be repeated?

9) The call letters of a radio station must have four letters. The first letter must be a K or a W. How many different call letter combinations are possible if letters may be repeated?

10) How many different four-digit ID tags can be made if repeats are allowed?

11) How many different four-digit ID tags can be made if it must start with a 7 and no repeats are allowed?

12) In how many different ways can the Harry Potter series of books (7 books total) be arranged in a row on a shelf?



13) In how many different ways can a manager select a pitcher - catcher combination if the manager has 5 pitchers and 2 catchers to choose from?

14) A coin is tossed 8 times. How many different outcomes are there for this series of 8 flips?

15) Six different colored tiles are available to make a pattern in a row of floor tile. How many possible different 4-color patterns are possible if no colors may be repeated?

16) Six different colors of tile are available to make a pattern in a row of floor tile. Many tiles of each color are available. How many 4-colored patterns can be made if colors may be repeated?

17) Four cards are dealt from a standard deck of 52 cards. In how many different orders of suit could the

cards be dealt? For example, one order is Club, Heart, Club, Diamond.

18) A pizza restaurant offers 6 different toppings for their pizzas. How many different pizzas are possible?



19) Use a tree diagram to find all possible outcomes for the result of a series of coin flips if the coin is flipped two times. Write a list of the possible results when complete.

20) The Super-Cool Ice Cream Shoppe sells sundaes, cones, or ice cream bars. You will pick either butterscotch or chocolate and you may choose to have it with nuts or without nuts.

a) Draw a tree diagram to illustrate the different types of ice cream treats that you could order.

b) How could you find the number of outcomes using the Fundamental Counting Principle?

c) How many different outcomes are possible?

21) A quiz has four true/false questions on it. Use a tree diagram to show all the different possible answer keys.

22) A box contains a $1 bill, a $5 bill, and a $10 bill. Two bills are selected one after the other without replacing the first bill. Draw a tree diagram to show all possible amounts of money that may be drawn.

23) The Eagles and Hawks play each other in a hockey tournament. The first team to win two games is the champion. Use a tree diagram to show all different possible outcomes for the tournament.

Review Exercises

24) Consider a situation in which a baseball manager must decide which one of four players will pitch (P1, P2, P3, or P4) and which one of 2 players will catch (C1 or C2).

a) What is the sample space for this situation?

b) How many outcomes are possible?

1.3 Permutations



Learning Objectives

- Know the definition of a permutation
- Be able to calculate the number of permutations using the permutations formula and with technology
- Understand the connection between the Fundamental Counting Principle and permutations

The Fundamental Counting Principle provides us with a tool that allows us to calculate the number of outcomes possible in many situations. What if the situation is a bit more complex? For many situations, the order that we complete a task does not matter. Ordering milk, bacon, and scrambled eggs in that order is the same as ordering bacon, scrambled eggs, and milk. In this case the order that we make our choices wouldn't matter, but there are many situations in which the order that we do things does make a difference.

A **permutation** is a specific order or arrangement of a set of objects or items. What if you wish to call someone on the phone? If I make the call, the order that I punch in the numbers matters so this is an example of a permutation. A good question to ask when deciding if your arrangement is a permutation is **"DOES ORDER MATTER?"** If yes, then you are dealing with a permutation. For example, if you ordered an ice cream sundae and they put the cherry in first, then the chocolate sauce, and then the ice cream, you would probably would not be happy with that particular ice cream sundae. You would likely prefer that they put the ice cream in first, then the chocolate sauce, and then put the cherry on top. Clearly each sundae had the same three ingredients, but they were quite different from one another. Each order that we can make the ice cream sundae is called a permutation.

There is a simple formula for figuring out how many permutations exist when 'r' objects are selected from a set of 'n' objects. The left side of the equation can be read "n P r", just as it looks or "n Permutations of size r".

$$_nP_r = \frac{n!}{(n-r)!}$$

Recall that the exclamation point is a factorial. For example, $5! = 5 \times 4 \times 3 \times 2 \times 1$. Also, be sure to find the permutations command on your calculator.

In our ice cream sundae discussion, 'n' would be 3 because there are 3 items to select from and 'r' would also be 3 because we are going to select all three items. Using the permutations formula, this would be ${}_3P_3 = \frac{3!}{(3-3)!} = \frac{3!}{0!} = \frac{6}{1} = 6$. In other words, there are 6 different orders that the ice cream sundae could be made. Note that 0! is equal to 1.

Example 1

Suppose you are going to order an ice cream cone with two different flavored scoops. You are going to take a picture of your ice cream cone for use in the school newspaper. The ice cream shop has 5 flavors to choose from; chocolate, vanilla, orange, strawberry, and mint. How many different ice cream cone photos are possible?

Solution

The first question to ask is "**Does Order Matter?**". If it does, then we are dealing with a permutation question. In this case, the order does make a difference. A chocolate on top of vanilla cone looks different than a vanilla on top of chocolate cone. We have five flavors to pick from, so n=5. We are going to select 2 flavors so r=2. ${}_5P_2 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{120}{6} = 20$ There 20 different permutations of ice cream cones we could order. The notation representing this situation, ${}_5P_2$, can be read as "Five 'P' Two" or "Five permutations of size Two". Be sure to perform this calculation using your calculator as well.

In the example above, you could have also found your answer using the Fundamental Counting Principle. There were 5 choices for the 1st flavor and then only 4 choices for the 2nd flavor. There are $5 \times 4 = 20$ ice cream cones possible.



Example 2

Give the value of ${}_6P_3$ by using the formula for permutations. Verify your solution on your calculator.

Solution

$${}_6P_3 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = \frac{720}{6} = 120$$

Example 3

Decide whether each of the situations below involves permutations.

a) A five-card poker hand is dealt from a deck of cards.

b) A cashier must give 3 pennies, 2 dimes, a 5 dollar bill, and a 10 dollar bill back as change for a purchase.

c) A student is going to open a padlock that has a three number combination.

d) A child has red, blue, green, yellow, and orange color crayons and will be coloring a rainbow using each color one time.

Solution

a) The order you get your five cards for a poker hand does not matter. If one of your cards was the ace of spades, it didn't matter if it was the first card or the last card dealt.

b) The order that the cashier gives you $15.23 in change does not matter as long as the total is $15.23.

c) The order you put in the three numbers for the combination makes a difference. If the correct combination is 12-27-19, the padlock will not open if you enter 19-12-27 even though the same three numbers are used.

d) The order that the child colors the rainbow does make a difference. The color pattern red, blue, green, orange, yellow will look different than green, blue, red, yellow, orange.

Problem Set 1.3

Exercises

1) Use the formula for Permutations, ${}_nP_r = \frac{n!}{(n-r)!}$ to find the value for each expression. Confirm each result by using your calculator.

a) ${}_8P_3$

b) ${}_4P_4$

c) ${}_5P_3$

d) ${}_5P_0$

2) How many 4 letter permutations can be formed from the letters in word *rhombus*?

3) For a board of directors composed of eight people, in how many ways can a president, vice president, and treasurer be selected?

4) How many different ID cards can be made if there are six digits on a card and no digit can be used more than once?

5) In how many ways can seven different types of laundry soap be displayed on a shelf in a store?

6) A child has four different stickers that can be placed on a model car in a vertical stack. In how many ways can this be done if each sticker is to be used only one time?

7) An inspector must select three tests to perform in a certain order on a manufactured part. He has a choice of seven tests. How many different ways can he perform three tests?

8) In how many different ways can 4 raffle tickets be selected from 50 tickets if each ticket wins a different prize?



9) A researcher has 5 different antibiotics to test on 5 different rats. Each rat will receive exactly one antibiotic and no rat will receive the same antibiotic as any other rat. In how many different ways can the researcher administer the antibiotics?

10) There are five violinists in an orchestra. Three of them will be selected to play in a trio with a different part for each musician. In how many ways can the trio be selected?

11) There are five violinists in an orchestra. Four of them will be selected to play in a quartet with a different part for each musician. In how many ways can the quartet be selected?

12) There are five violinists in an orchestra. All five of them will be selected to play in a quintet with a different part for each musician. In how many ways can the quintet be selected?

13) There are five violinists in an orchestra. A piece of music is written so that it can be played with either 3, 4, or 5 violinists. Each musician selected to play this piece will play a different part. In how many ways can a group of at least three musicians be selected? Hint: Use your answers from problems 10), 11) and 12).

14) Decide whether each situation below involves permutations. Briefly explain your answers.

a) Sophia picks three color crayons from a box of 12 crayons to make a picture for her cat Butterscotch.

b) A five-digit code is needed to open up an electronic lock on a car.

c) Twenty race car drivers must each complete three laps at a race track during a time trial, one after another, in order to establish the order in which the cars will start a race the next day.

d) There are seven steps that a student must follow when preparing cookies during their Family and Consumer Sciences course.

Review Exercises

15) Use the Fundamental Counting Principle to determine the number of different ways a person could order a meal if they are to pick one entree from four choices, one side order from three choices, and one drink from four choices.

16) A student wishes to check out three books from the library. She will check out one historical fiction book, one biography, and one book on art history. Build a tree diagram to show how many ways can this be done if there are two historical fiction books, three biographies, and two books on art history that she is considering checking out.

17) How many different outcomes are possible for the total on a roll of two dice if one die has 6 sides and one die has 4 sides?

1.4 Combinations



Learning Objectives

- Know the definition of a combination
- Be able to calculate the number of combinations using the combinations formula and with technology

We just looked at situations in which order matters. What if order does not matter? Suppose you have a younger brother or sister and your family goes out to a restaurant. There is a children's menu with activities at the restaurant that all the kids get. The owner of the restaurant has decided that each child will receive two different colored crayons to use on their menu. The restaurant happens to carry five colors of crayons: orange, yellow, blue, green, and red. This is a situation in which the order that the child gets their two color crayons does not matter. If you gave a child a red crayon and then a blue crayon, it would be the same as if you gave the child a blue crayon followed by a red crayon. As with permutations, the first question to ask is **"Does Order Matter?"**. When the order does not matter, you are dealing with a situation that involves **combinations**.

Example 1

Consider the color crayon problem in the previous situation. Make a list showing all of the different color crayon combinations that might occur. Be organized so as not to repeat any combinations.



Solution

To be organized, use the letters O, Y, B, G, and R to represent the five colors (Orange, Yellow, Blue, Green, and Red). Alphabetizing the list to insure that we don't skip any combinations gives us BG, BO, BR, BY, GO, GR, GY, OR, OY, RY Notice that while we have BG, we don't have GB as that would be a repeat. It appears that there are 10 combinations possible.

As with the Fundamental Counting Principle, we now must ask the question "How can we find the solution quickly?" Making a list works nice, but it could get a bit messy if the restaurant had 24 colors to choose from instead of 5 because our list would get very long. Out of curiosity, you may have tried ${}_5P_2$. However, when you work this out, you find that this gives us a result of 20 instead of 10. We must modify this formula for situations involving combinations.

Shown below is the formula for finding how many combinations are possible when order ***does not*** matter.

$$ {}_nC_r = \frac{n!}{r!(n-r)!} $$

As with the permutation formula, the 'n' stands for the number of objects available and the 'r' stands for the number of objects that will be selected.

Example 2

Consider the color crayon problem once again. Use the formula to find out the number of different color crayon combinations that are possible.

Solution

In our problem, 'n' is equal to 5 and 'r' is equal to 2. Our calculation would be ${}_5C_2 = \frac{5!}{2!(5-2)!} = \frac{5!}{2!\cdot 3!} = \frac{120}{2\cdot 6} = \frac{120}{12} = 10$. Be very careful that you find the result for the denominator before you divide! Find the ${}_nC_r$ command on your calculator and verify that ${}_5C_2$ is indeed equal to 10.

Example 3

Suppose that there are 12 employees in an office. The boss needs to select 4 of the employees to go on a business trip to California. In how many ways can she do this?

Solution

We first ask whether the order that the employees are selected matters. In this case, the answer is no because either you will be going on the trip or you won't be going. Being the fourth name on the list of people who get to go is just as good as being the first name on the list. We have 12 people to select from and we will be selecting 4 or ${}_{12}C_4 = \frac{12!}{4!\cdot(12-4)!} = 495$. There are 495 possible combination of groups of 4 that might be selected to go on the trip to California.

Problem Set 1.4

Exercises

1) Use the formula for combinations to find the value of each expression. Use a calculator to verify each answer.

a) ${}_5C_5$

b) ${}_6C_4$

c) ${}_3C_0$

d) ${}_7C_3$

2) In how many ways can 3 cards be selected from a standard deck of 52 cards?

3) In how many ways can three bracelets be selected from a box of ten bracelets?

4) In how many ways can a student select five questions to answer from an exam containing nine questions?



5) In how many ways can a student select five questions to answer from an exam containing nine questions if the student is required to answer the first and the last question?

6) The general manager of a fast-food restaurant chain must select 6 restaurants from 11 for a promotional program. In how many different possible ways can this selection be done?

7) There are 7 women and 5 men in a department. In how many ways can a committee of 4 people be selected?

8) For a fundraiser, a travel agency has donated 5 free vacations to Mexico as grand prizes in a raffle. Suppose that 220 people paid for raffle tickets. In how many different ways can the vacation winners be selected?

9) A high school choir has 27 female and 19 male members. Two students will be selected from the choir to represent the school in the All-State Choir.

a) In how many ways can the director select two students if she decides both students will be female?

b) In how many ways can the director select two students if she decides both students will be male?

c) In how many ways can the director select two students?

d) Using your answers from a), b) and c), determine how many ways the choir director can select two students such that one student will be a male and one student will be a female?

Review Exercises

10) In how many ways can the team captain of a kickball team arrange the kicking order for the 7 players on the team?

11) An electronic car door lock has five buttons on it and each button has a different digit. Suppose the combination to unlock the door is 4 digits long.

a) How many different combinations are possible if digits can be repeated?

b) How many different combinations are possible if digits cannot be repeated?

12) Give the sample space for the different results that may occur if a coin is flipped twice.



13) Decide whether each situation involves permutations.

a) A teacher must pick two students from a class of 30 to put their answers on the board for problem #11 from last night's homework.

b) In order to be allowed outside to play in the rain, a 5-year old must put on socks, shoes, and boots.

c) A student has a strict bedtime of 11 pm. They need two hours to finish writing a paper, one hour for a math assignment, two hours for a science experiment, two hours to practice piano, and 1 hour for relaxation. It is 6 pm right now.

1.5 Mixed Combinations and Permutations



Learning Objectives

- Determine whether a situation involves permutations or combinations
- Understand the mathematical implications of the words 'and' & 'or'

Having covered the basics of combinations and permutations, you are ready to have a mixture of problems with slight variations. A common variation involves an understanding of some key words used in

mathematics. Commonly, the word "**and**" indicates multiplication and the word "**or**" indicates addition. Consider the examples below.

Example 1

In how many ways can committee of 3 people be chosen if there are 8 men and 4 women available for selection and we require that two men and one woman be on the committee?

Solution

The order that we place the people on a committee does not matter. It makes no difference if you are the first person or the last person selected for the committee. Either you are on the committee or you are not on the committee, therefore this is a combination question. Notice that we want two men **and** one women. The word 'and' indicates multiplication. In other words, we will look for the product of how many ways we can select two men from eight and one women from four. ${}_8C_2 \times_4 C_1 = 28 \times 4 = 112$. There are 112 ways to select this committee of 3 people.

Example 2

In how many ways can a committee of 5 people be chosen if there are 7 men and 5 women available for selection and we require *at least* 4 women on the committee?

Solution

We first ask "Does order matter?". In this case, the order that someone is placed on a committee does not matter. Either you are on the committee or you are not. Once again, we are dealing with a combination question. The key phrase in this example is *at least*. This can be interpreted to mean that we either select 4 women **and** 1 man **or** 5 women **and** 0 men. Remember that the word 'and' indicates multiplication and the word 'or' indicates addition. It looks like we are going to have some addition and multiplication in this problem.

${}_5C_4 \times_7 C_1 +_5 C_5 \times_7 C_0 = 5 \times 7 + 1 \times 1 = 35 + 1 = 36$. There are 36 ways to put this committee together.

Example 3

In a certain country, there are two political parties. Each party is responsible for nominating both a presidential and vice-presidential candidate. The candidates will participate in a debate once they are chosen. In the first party, there are 6 candidates available and in the second party there are 5 candidates available. How many different debate combinations are possible?

Solution

The order that we select the candidates does make a difference. Selecting party member 'A' for a presidential candidate and party member 'B' for a vice-presidential candidate is different

than selecting party member 'B' for a presidential candidate and party member 'A' for a vice-presidential candidate. Therefore, this is a permutations question. Since we will select candidates from the first party **and** candidates from the second party, we expect there to be multiplication in this problem as well.

${}_6P_2 \times {}_5P_2 = 30 \times 20 = 600$. There are 600 different ways that the debate participants can be chosen.

Problem Set 1.5

Exercises

1) In your own words, state how you can tell the difference between a combination and permutation problem.

2) Your closet contains 10 different styles of shoes. In how many ways can you pick out five different styles of shoes for the school week if you don't care which day of the week you wear each style?

3) Your closet contains 10 different styles of shoes. In how many ways can you pick out five different styles of shoes for the school week if you do care which day of the week you wear each style?

4) You are drawing a rainbow using five different colored crayons from your box of 24 colors. In how many ways can you draw a rainbow if the first color you pick will be the top layer and so on?



5) You are drawing a rainbow using five different colored crayons from your box of 24 colors. In how many ways can you pick the five colors for your rainbow?

6) Suppose 5 cards are dealt from a standard deck of 52 cards.

a) How many unique 5-card hands are possible?

b) In how many different orders can those 5 cards be dealt?

7) Suppose the majority party in a foreign country must select a prime minister and secretary of state from an eligible group of 36 party members. In how many ways can this be done?

8) There are 7 women and 5 men in a department. Four people are needed for a committee.

a) In how many ways can a committee of 4 people be selected?

b) In how many ways can this committee be selected if there must be *exactly* 2 men and 2 women on the committee?

c) In how many ways can this committee be selected if there must be *at least* 2 women on the committee?

9) A company has 8 cars and 11 trucks. The state inspector will select 3 cars and 4 trucks to be tested for safety inspections. In how many ways can this be done?

10) In a train yard there are 4 tanker cars, 12 boxcars, and 7 flatcars available for a train. In how many ways can a train be made up consisting of 2 tanker cars, 5 boxcars, and 3 flatcars?

11) Flakes-R-Us cereal comes in two types, Sugar Sweet and Touch O' Honey. If a researcher has ten boxes of each type, how many ways can she select two boxes of each for a quality control test?

12) In how many ways can a jury of 12 people be selected from a pool of 12 men and 10 women?



13) In how many ways can a jury of 6 men and 6 women be selected from a pool of 12 men and 10 women?

14) A corporation president must select a manager and assistant manager for each of two stores. In how many ways can this be done if the first store has 9 employees and the second store has 7 employees? (Employees will stay at their current stores.)

15) Suppose that in this trimester that every sophomore is required to take 2 math classes, 2 social studies classes, and a reading class. How many different combinations of teachers are possible for a given student if there are 9 math teachers, 12 social studies teachers, and 4 reading teachers available? (No student will have the same teacher for two different hours.)

16) In how many different ways can six people be assigned to three offices if there will be two people in each office?

Review Exercises

17) Use the formula for combinations to find the value of ${}_7C_3$.

18) Use the formula for permutations to find the value of ${}_6P_4$.

19) In how many ways can the letters in the word 'magic' be arranged?

20) How many different outcomes are possible for the total of when two 4-sided dice are rolled?

21) A teacher will select three students to work problems on the board from her class of 34 students. In how many ways can this be done if the three problems to be worked are #11, #14, and #26?

1.6 Chapter 1 Review



There are three primary counting methods that are commonly used in probability: the Fundamental Counting Principle, combinations, and permutations. The Fundamental Counting Principle states that to find the number of outcomes for a given situation, simply multiply the number of ways each event may occur by each other. When deciding whether to use combinations or permutations, you must ask if the order matters. If so, use permutations, otherwise use combinations. When working with counting, make sure that you have an organizational strategy. For example, use an alphabetized list, a sketch, or a tree diagram. This will make it much easier for you to come up with the sample space.

Chapter 1 Review Exercises

1) Suppose that two 5-sided dice are rolled.

a) Draw a grid showing all the outcomes for the different totals that may occur.

b) Use {brackets} to write down the sample space.

c) Suppose a friend offers to play a game in which you are paid \$4 any time a number divisible by 4 occurs. Otherwise you pay your friend \$2. If you decide to play, would you expect to win money or lose money? Use your grid from part a) to help explain your answer.

2) The lunch at The Diner has a choice of ham, turkey, or roast beef on rye or white bread with coffee or milk. Draw a tree diagram that illustrates what a person might have for lunch if they pick only one meat, one bread, and one drink.

3) Find the value for each expression below. Show your work by hand and use your calculator to verify your results.

a) 5!

b) ${}_6P_3$

c) ${}_7C_5$

d) (5 – 2)!

e) 4! – 2!

4) There are four runners in a race. In how many ways can the runners finish the race?

5) A store has eighteen outfits available for display, but only six outfits can be used for a window display. If each order counts as a new arrangement, how many different arrangements are possible?

6) Paul has three baseballs and four bats. How many possible ball and bat combinations can he choose?

7) How many license plates are possible if each plate must have three letters followed by three digits (repeats are allowed)?



8) How many license plates are possible if each must have three letters followed by three digits (repeats are not allowed)?

9) There are twenty candidates in the Mr. Minnesota contest. How many ways could the judges choose the winner, first-runner up, and second-runner up?

10) The yearbook editor must select two photos out of 42 juniors and two out of the 45 seniors for a page in the yearbook. How many photo combinations are possible?

11) A homeless shelter has decided to purchase all new kitchen appliances. They need one oven, one refrigerator, and one dishwasher. The appliance store has 7 brands of ovens, 6 brands of refrigerators, and 5 brands of dishwashers. In how many brand arrangements can they purchase their appliances?

12) An ice cream shop has 8 different flavors of ice cream available. How many 2-scoop cones can be made if you are allowed to have the same flavor for both scoops?

13) An ice cream shop has 8 different flavors of ice cream available. How many 2-scoop cones can be made if you decide not to have the same flavor for both scoops?

14) Suppose a jury of 12 is being selected from a pool of 20 candidates. In how many ways can this be done?

15) Suppose a jury of 12 is being selected from a pool of 13 men and 7 women. In how many ways can this be done if the judge states that the jury must contain *exactly* 5 women?

16) Suppose a jury of 12 is being selected from a pool of 13 men and 7 women. In how many ways can this be done if the judge states that the jury must contain *at least* 5 women?

17) In how many ways can I put together an outfit if I have 7 shirts, 5 pairs of pants, and 4 hats from which to choose?

18) For $7.99, a restaurant will sell you their lunch special. The special is either a hamburger or chicken sandwich, onion rings or fries, and soda or coffee.

a) Make a tree diagram showing the different ways a customer may order the lunch special.

b) How many outcomes are there? Use the Fundamental Counting Principle to justify your answer.

Image References

Breakfast http://worldaffairspittsburgh.blogspot.com

Tetrahedral Dice http://www.bbc.co.uk

Harry Potter Books http://www.dipity.com

Ice Cream Cones http://www.bunrab.com

Raffle Ticket http://canuckamusements.com

Color Crayon http://www.rosespet.com

Exam http://www.iphlebotomycertification.com

Coin http://www.marshu.com

Rainbow http://tracynicholls.webs.com/

Jury http://adriandayton.com

License Plate http://www.15q.net

Chapter 2

Calculating Probabilities

2.1 Calculating Basic Probabilities

Learning Objectives

- Understand how to calculate and write a probability
- Understand what constitutes chance behavior
- Understand the concept of the Law of Large Numbers

Probabilities give us an idea of how likely it is for a certain event to happen. For example, when a coin is flipped, the chance that it comes up heads is 50%. Probabilities can be expressed as decimals, fractions, percents, or ratios. We could have said the probability of flipping heads is , 0.5, $\frac{1}{2}$, 50% or 1:2. Each of these conveys the idea that we should expect to get a heads half of the time. Probabilities only give us an idea of what to expect in the long run. However, they do not tell us what will happen in the short term.



Suppose we flip a coin 10 times in a row and get heads every single time. The next coin flip is still a **random event** because while we cannot tell for certain what the next flip will be, we can be certain that about 50% of all tosses over a long set of tosses will be heads. Some people think that we are on a roll so we are more likely to get another heads. Others will say that getting tails is more likely because we are due to get tails. The truth is that we cannot tell what will happen on the next flip. The only thing we know for certain is that there is a 50% chance that the coin will be heads on its next flip. If we continue to flip this same coin hundreds of times, we would expect the percent of heads to get closer and closer to 50%.

Chance Behavior is not predictable in the short term, however, it has long term predictability. The **Law of Large Numbers** tells us that despite the results on a small number of flips, we will eventually get closer to the **theoretical probability**. The theoretical probability should match what would happen in the long run for some random event. The outcomes in any random event will always get close to the theoretical probability if the event is repeated a large number of times. We might roll a dice 4 times in a row and get a 6 each time, however, if we rolled this dice hundreds of times, the percent of time that we get a 6 gets closer and closer to the theoretical probability of $\frac{1}{6}$.

When calculating a probability, we divide the number of favorable outcomes (outcomes we are interested in) by the total number of outcomes. In other words, the probability that outcome 'A' occurs is found by the formula $P(A) = \frac{\#\ of\ favorable\ outcomes}{total\ \#\ of\ outcomes}$.



Consider a standard deck of 52 playing cards.

Standard Deck of 52 Playing Cards

Clubs	Spades	Hearts	Diamonds
A♣	A♠	A♥	A♦
2♣	2♠	2♥	2♦
3♣	3♠	3♥	3♦
4♣	4♠	4♥	4♦
5♣	5♠	5♥	5♦
6♣	6♠	6♥	6♦
7♣	7♠	7♥	7♦
8♣	8♠	8♥	8♦
9♣	9♠	9♥	9♦
10♣	10♠	10♥	10♦
Jack♣	Jack♠	Jack♥	Jack♦
Queen♣	Queen♠	Queen♥	Queen♦
King♣	King♠	King♥	King♦

If we asked the question "What is the probability of being dealt a face card (jack, queen, or king)?", we would need to count how many cards are face cards and then divide by the total number of cards 52.

In this situation there are 12 face cards and 52 cards overall so our probability of getting a face card is $\frac{12}{52} = \frac{3}{13} \approx 0.23$.

In probability, there are outcomes that are sure to happen and there are outcomes that are impossible. If we are once again dealing with a standard 52 card deck, the chance of being dealt either a red card or a black card if one card is dealt is 100%. The chance of being dealt a blue card is 0% since there are no blue cards in a standard deck. All random events have probabilities between 0 and 1. In addition, the sum total of the probabilities for all possible outcomes in the sample space is equal to 1. In other words, if an event occurs, there is a 100% chance that one of the possible outcomes will happen. The list below summarizes these rules.

a) The probability of a sure thing is 1.

b) The probability of an impossible outcome is 0.

c) The sum of the probabilities of all possible outcomes is 1.

d) The probability for any random event must be somewhere from 0 to 1.

As shown earlier, we notate the probability of event 'A' happening as P(A). For example, the probability of rolling a three on a six-sided die can be written $P(3) = \frac{1}{6}$. Sometimes we are interested in the probability of an event not occurring. This is called the **complement** of the event. We can write the probability of the complement of event 'A' happening as P(˜A), P(not A), or $P(A^c)$. The formula for the complement of an event is $P(\textit{not } A) = 1 - P(A)$. On our die rolling question, P(˜3)= $1 - P(3) = 1 - \frac{1}{6} = \frac{5}{6}$. In other words, there is a $\frac{5}{6}$ chance of the dice not landing on a 3. It is important to notice that the probability of an event happening and the probability of its complement always add up to 1.



Example 1

Which of the following situations are random events?

i) A student looks through their closet to decide what shirt to wear to school.

ii) A student labels each of their 6 pairs of shoes 1 through 6 and then rolls a single die to decide which pair to wear.

iii) The state legislature decides to increase funding to schools by 3%.

iv) A professional golfer makes a hole-in-one on a 200 yard hole.

Solution

Situations i) and ii) are not random events. In both cases, there are additional factors that are influencing the decision. The day of the week or the temperature outside might influence your shirt choice and how much money the state legislature happens to have might influence funding.

Both situations iii) and iv) are random events because while we can't predict what will happen in this particular instance, we can make long term predictions. We can predict the percent of the time the student might end up with the shoes labeled #2 and we can predict the percent of the time that the golfer will make a hole-in-one based upon previous performance.

Example 2

In the game of pool, there are a total of 15 balls. Balls numbered 1-8 are solid and balls 9-15 are striped. There is only one black ball (the eight ball). There are two of every other color of pool ball.



Suppose the pool balls were put in a bag and a single pool ball is pulled out of the bag. What is the probability that the ball:

a) is yellow?

b) is striped?

c) has a number on it that is greater than 10?

d) is not striped?

Solution

a) $P(Y) = \frac{2}{15} \approx 0.13$

b) $P(Striped) = \frac{7}{15} \approx 0.47$

c) $P(> 10) = \frac{5}{15} = \frac{1}{3} \approx 0.33$

d) P(˜*Striped*)=1 − *P*(*Striped*) = $1 - \frac{7}{15} = \frac{8}{15} \approx 0.53$

In addition to these types of questions, we can also calculate probabilities by incorporating our counting methods from Chapter 1. Recall that the probability of an event occurring is the number of favorable outcomes divided by the number of total possible outcomes.

Example 3



A jury of 12 people is to be selected from a group of 12 men and 8 women. What is the probability that the jury has at least 6 women on it?

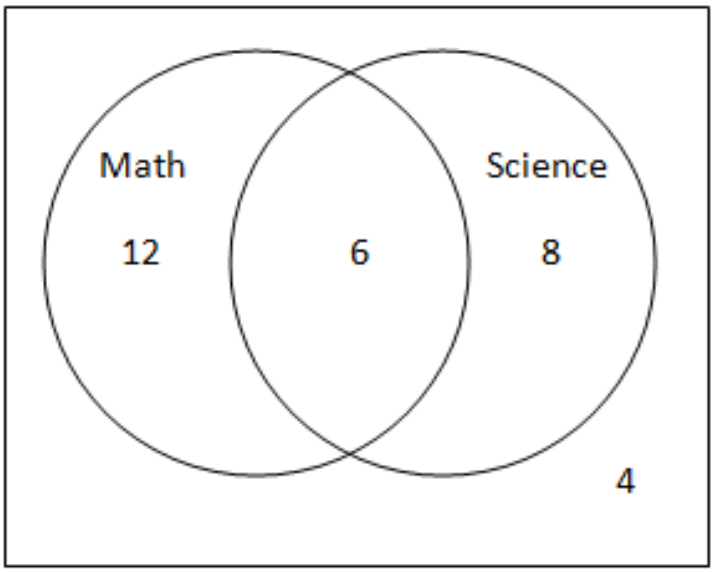
Solution

The total number of outcomes possible is based upon selecting 12 members from a pool of 20. Since order will not matter, there are ${}_{20}C_{12} = 125,970$ ways to pick a jury of 12. We now want to have at least 6 women on the jury. This means we could have 6 women and 6 men **or** 7 women and 5 men **or** 8 women and 4 men on the jury. Mathematically, this would be ${}_8C_6 \times {}_{12}C_6 + {}_8C_7 \times {}_{12}C_5 + {}_8C_8 \times {}_{12}C_4 = 28 \times 924 + 8 \times 792 + 1 \times 495 = 25,872 + 6,336 + 495 = 32,703$. There are 32,703 ways to have at least 6 women on the jury out of a possible total of 125,970 different juries or $\frac{32,703}{125,970} \approx 0.26$. There is about a 26% chance that the jury will have at least 6 women on it.

Sometimes, data is organized in a **Venn Diagram**, as shown in the example below. We will examine these in greater depth in section 2.3 but for now, it is important to understand that a Venn Diagram is an organizational tool that makes it easier to interpret a situation and answer basic probability questions.

Example 4

A class of 30 students is surveyed to see whether or not they had a science class and/or a math class this trimester. There are 18 students that have a math class, 14 students who have a science class, and 4 students who have neither. It also turns out that this includes 6 students who currently have both classes. The results of the survey are shown in the Venn Diagram below.



a) How many total students are taking a math class this trimester?

b) What is the probability that a randomly selected student is taking a math class this trimester?

c) What is the probability that a randomly selected student is taking both a math and science class this trimester?

d) What is the probability that a randomly selected student is not taking either a math or science class this trimester?

Solution

a) There are 12 kids who only have a math class and 6 kids who have both a math a science class this trimester for a total of 18 kids.

b) $P(Math) = \frac{18}{30} = \frac{3}{5} = 0.6$

c) $P(Math \ \& \ Science) = \frac{6}{30} = \frac{1}{5} = 0.2$

d) $P(NoMathorScience) = \frac{4}{30} = \frac{2}{15} \approx 0.13$

Problem Set 2.1

Exercises

For problems 1-5, express your answer both as a fraction (reduce if possible) and as a decimal to the nearest hundredth.

1) Suppose a single card is dealt from a standard deck of 52 cards. Find the probability that the card is:

a) a red card.

b) a face card.

c) an ace.

d) a three.

e) a club.

f) the three of clubs.

g) a black king.

h) not a spade.

2) A bag contains some jelly beans. There are a total of 6 red jelly beans, 4 green jelly beans, 2 black jelly beans, 5 yellow jelly beans, and 3 orange jelly beans in the bag. Suppose one jelly bean is drawn from the bag.

a) Find P(purple).

b) Find P(yellow).

c) Find P(˜red).

3) The game Scattegories® uses a 20-sided die. It has all the letters of the alphabet on it except Q, U, V, X, Y, and Z. Find each probability below if the die is rolled one time.

a) P(Vowel)

b) P(˜Vowel)

c) P(Q)

d) P(Q^c)

e) P(a letter alphabetically after Q)



4) A single 6-sided die is rolled one time. Find the probability that the result is:

a) a three

b) a seven

c) an even number

d) a prime number

e) a number equal to or greater than 5.

5) The month of October in a 2011 calendar has 31 days with October 1st being a Saturday. Suppose a day is randomly selected. Find each probability.

a) P(weekend)

b) P(not a weekend)

c) P(October 31st)

d) P(October 32nd)

e) P(˜October 31st)

f) P(an odd-numbered day)



October 2011 printablecalendars.resources2u.com

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

6) A roulette wheel contains 38 slots. When the wheel is spun, a ball is dropped onto the wheel and the ball will stop on one of the slots. There are 18 black slots, 18 red slots, and 2 green slots. Suppose the ball on a roulette wheel has landed on red four times in a row. What is the chance that the ball will drop on red on the next spin?



7) A coin has been flipped 10 times. Suppose that it has come up heads on only 2 out of those ten times.

a) What percent of the time has the coin come up heads?

b) Suppose we flip the coin 90 more times and 45 of those 90 flips come up heads. Of the 100 flips completed so far, what percent of the time has the coin come up heads?

c) Suppose we continue to flip the coin an additional 900 times and that 450 of those 900 flips come up heads. Of the 1000 flips completed, what percent of the time has the coin come up heads?

d) As we flipped the coin more and more, the percentage of heads got closer and closer to 50% despite the fact that only 2 of the first 10 flips were heads. What rule does this illustrate?

8) Two 6-sided dice are rolled and we keep track of the total on the two dice.

a) Make a 6 by 6 grid showing the different totals that you can get when rolling the two dice.

b) What is the probability that you get doubles?

c) What is the probability that you get a total of 7?

d) What is the probability that you get a total of at least 8?



9) The high school concert choir has 7 boys and 15 girls. The teacher needs to pick three soloists for the next concert but all of the members are so good she decides to randomly select the three students for the solos.

a) In how many ways can the teacher select the 3 students?

b) What is the probability that all three students selected are girls?

c) What is the probability that at least one boy is selected?

10) A test begins with 5 multiple choice questions with four options on each question. It then has 5 true/false questions.

a) How many answer keys are possible?

b) What is the probability of getting every question correct if a student guesses on each question. Leave your answer as a fraction.

11) A lawn and garden store is moving locations and needs to move its riding lawn mowers to the new store. They have 8 mowers with 36-inch decks, 15 mowers with 42-inch decks, and 6 mowers with 48-inch decks that need to be moved. The trailer they are using can move a total of 8 mowers on each load so several trips will have to be made.

a) In how many ways can 8 mowers be randomly selected for the first load?

b) What is the probability that all the mowers with 48-inch decks get selected for the first load? Leave your answer as a reduced fraction.

c) What is the probability that the first load has exactly two 36-inch deck mowers, four 42-inch deck mowers, and two 48-inch deck mowers?



Review Exercises

12) In how many ways can three students be selected for a committee if there are 11 students from which to select?

13) A hockey player needs new skates, a new helmet, and a new stick. Hockey Central has 5 brands of skates, 6 brands of helmets, and 8 brands of sticks. In how many different ways can the player select one of each item?

14) Two standard 6-sided dice are rolled and the results from the two dice are added together. Build a grid to determine which outcome is most likely to occur.

15) On a TV game show, three contestants must each pick a box which they believe contains the grand prize based upon clues given about each box. In how many different ways can this be done if there are 10 boxes from which to choose?

2.2 Compound and Independent Events



- Understand how to perform the calculations for compound events
- Compute probabilities for situations with and without replacement
- Understand when two events are independent
- Understand how to compute the probability when two independent events occur

From section 2.1, you found that it is quite straightforward to calculate probabilities for simple situations. What happens when we calculate probabilities from multiple events? For example, suppose you roll a single die and then flip a coin. What are the chances that the die comes up with a 5 and the coin gives you a heads? A situation that asks you to calculate probabilities for a situation that involves two or more events or steps is called a **compound event.** We will try to find out how to handle these types of situations by examining several situations and then making a conclusion.

Example 1

Suppose a single die is rolled and a coin is flipped. What is the probability that the die comes up with a 5 and the coin gives you a heads? Use a list to help you find out.

Solution

Start with a list of all the possible outcomes. 1H, 1T, 2H, 2T, 3H, 3T, 4H, 4T, 5H, 5T, 6H, 6T. There are 12 equally-likely outcomes. Of these, only 5H is a five with a heads. Therefore, the answer is $\frac{1}{12}$.

What you might have noticed is that P(five)=$\frac{1}{6}$ and P(heads)=$\frac{1}{2}$. Curiously, $\frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$. Is this just a coincidence or is there something more here? You might recognize that flipping a coin does not effect what you roll on the die. When two events do not have an impact on each other, the events are called **independent**. Consider the two situations below.

Situation 1: Suppose your teacher picks students to do problems on the board. After each student does their problem, the teacher gives the student a piece of candy. Because your teacher wants make sure that every student gets a chance to do a problem and get a piece of candy, she keeps track of who has worked problems on the board. The selection of the next student is not independent of previous selections the teacher has made.



Situation 2: Does a coin have a memory? As far as we can tell, the answer is no. This suggests then that a coin does not pay attention to whether it came up heads or tails. It does not want to make sure that the same number of heads come up as tails. Even if it comes up heads many times in a row, the next flip of that coin is not influenced whatsoever by the previous flips. Successive coin flips are independent of one another.

Example 2

Decide which pairs of events below are independent.

i) Two cards are dealt, one after the other, from a standard deck of 52 cards.

ii) A spinner with three colors is spun twice.

iii) A single die is rolled and a coin is flipped.

iv) You play on the school baseball team and you win a carnival game by throwing a baseball to try to break a plate.

Solution

Situations ii) and iii) represent pairs of independent events. The result from the first spin of the spinner does not affect the result of the second spin of the spinner. The result of the roll does not impact what happens when the coin is flipped.

Situations i) and iv) are not independent. Once the first card from the deck is dealt, the probabilities for what the second card might be will change. For example, if the first card was the ace of spades, it is impossible for the second card to also be the ace of spades. Being a baseball player makes it more likely mean that you are accurate and can throw a ball harder than a typical person and you therefore would be more likely to break a plate.

Let's do another example involving calculations and investigate if multiplying probabilities in a situation involving independent events gives us the correct result.

Example 3

A coin is flipped three times in a row. What is the probability that all three flips result in heads? Find your answer by either using a tree diagram or by making a list.

Solution

To be organized, we can make an alphabetically list. HHH, HHT, HTH, HTT, THH, THT, TTH, TTT. There are 8 outcomes altogether and only one of those is HHH. P(HHH)=$\frac{1}{8}$.

The probability of heads on one flip is $\frac{1}{2}$. As in example 1, we have P(HHH)=$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. Note that the three coin flips are independent of each other.

In both Example 1 and Example 3, we could multiply the probabilities of each individual event to get the probability of both events happening. This is always true of independent events. In other words, suppose we want the probability of both some outcome 'A' from one event and some outcome 'B' from a second event that is independent of the first event. If the probability of our first outcome is P(A) and the probability of our second outcome is P(B), then the probability of both A and B happening is P(A and B)=$P(A) \times P(B)$.

For Independent Events

$$P(A \,\&\, B) = P(A) \cdot P(B)$$

Example 4

You are dealt one card from each of two separate decks of cards.

a) What is the probability that both cards are the king of clubs?

b) What is the probability that the two cards are identical?

Solution

In both situations, the events are independent.

a) We have P(K♣ & K♣)=$\frac{1}{52} \times \frac{1}{52} = \frac{1}{2704} \approx 0.00037$.

b) There are two good ways to solve this problem. We could imagine that we do what we did in part a) for all 52 cards in the deck. We simply could multiply our answer for part a) by 52 to get $\frac{1}{52} \approx 0.019$. Another way to think about this would be to ask about each card separately. What is the chance that the first card could be useful in making a match? (100%) What is the chance that the second card will be useful in making a match? If we already have one card picked, the chance that the card from the second deck will match it is $\frac{1}{52}$. $1 \times \frac{1}{52} = \frac{1}{52} \approx 0.019$.

Multiplication of probabilities expands to more than just two independent events. It also works with three or more independent events and it even works with many situations that do not have independent events. In general, when finding the probability of compound events, multiply the probabilities of each

individual event. If we are interested in the probability of events A, B, and C happening, we can multiply $P(A) \times P(B) \times P(C)$.

This same principle also works with compound events in which we distinguish whether or not we have **replacement**. Suppose we are asked to pick two cards out of a deck. If we are asked to do this without replacement, we will select the first card and record what it is. When we select our second card, we must remember that the deck has changed. Nonetheless, we can find the probability of drawing these two particular cards by multiplying the individual probabilities.

Example 5

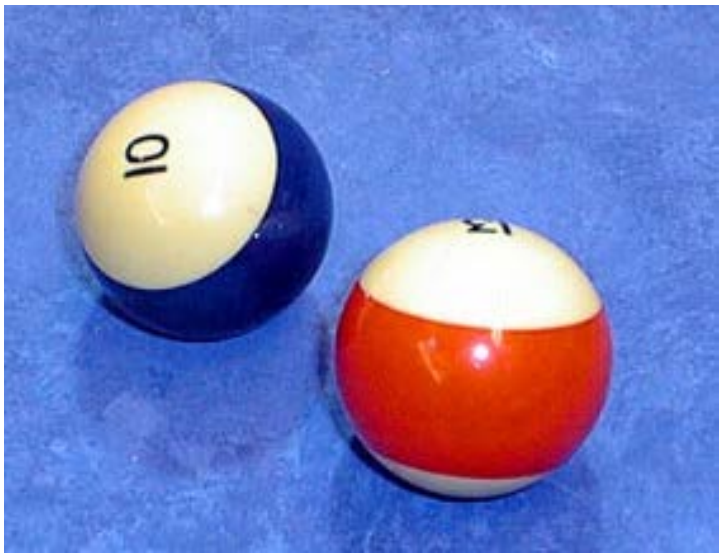
Suppose you have a set of pool balls in a bag. You pull two pool balls out of the bag, one after the other, *without* replacement. What is the probability that both pool balls are striped?

Solution

There are 7 striped pool balls out of the 15 pool balls. The chance that the first pool ball is striped is $\frac{7}{15}$. Since we are not going to replace the first pool ball, what is in the bag has now changed. There are only 14 pool balls left of which 6 are striped since the first one removed from the bag was also striped. The chance that the second pool ball is striped is $\frac{6}{14}$. To find the probability that both pool balls are striped, we multiply the individual probabilities. This gives $\frac{7}{15} \times \frac{6}{14} = \frac{42}{210} = \frac{1}{5} = 0.2$. There is a 20% chance that both balls will be striped if we use replacement.

Example 6

Suppose you have a set of pool balls in a bag. You pull two pool balls out of the bag, one after the other, *with* replacement. (This means that after you record what the first ball is, you put it back into the bag and remix the pool balls before you select the second pool ball.) What is the probability that both pool balls are striped?



Solution

There are 7 striped pool balls out of the 15 pool balls. The chance that the first pool ball is striped is $\frac{7}{15}$. Since we are going to replace the first pool ball, what is in the bag has not changed. There are still 15 pool balls of which 7 still are striped. Therefore, the chance that the second pool ball is also striped is $\frac{7}{15}$. To find the probability that both pool balls are striped, we multiply the individual probabilities to get $\frac{7}{15} \times \frac{7}{15} = \frac{49}{225} \approx 0.22$. There is approximately a 22% chance that both balls will be striped if we do not use replacement.

We also run into situations where we are dealing with compound events involving very large populations. In these sorts of situations, we must be careful about how we interpret the mathematics.

Example 7

Approximately 20% of all Americans smoke. Suppose two Americans are selected at random. What is the probability that both Americans are smokers?

Solution

The chance that the first person is a smoker is 20%. Some students think that the chance that the second person is a smoker changes after the first person is selected, however, it does not. The population of America is so large that selecting a single person out from that population will not affect the overall percentage of Americans that smoke. The probability that the second person smokes is also 20%. P(2 smokers selected)=$0.2 \times 0.2 = 0.04 = 4\%$.

Example 8

Approximately 20% of all Americans smoke. Suppose five Americans are selected at random. What is the probability that all five are *non-smokers*?

Solution

Since 20% of Americans are smokers, 80% must be non-smokers. This gives us $0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = (0.8)^5 \approx 0.33$**.** The chance that all five Americans selected will be non-smokers is about 33%.

Problem Set 2.2

Exercises

1) What does it mean for two events to be independent?

2) Suppose you are dealt one card each from two separate decks of cards. What is the probability that both of your cards are:

a) red?

b) spades?

c) jacks?

d) face cards?

3) For each situation below, determine whether the two events are independent.

a) Flip a coin and then draw a card from a standard deck of 52 cards.

b) Draw a marble from a bag, do not replace it, and then draw a 2nd marble from the same bag.

c) Get a raise at work and purchase a new car.

d) Drive on ice and lose control of your car.

e) Have a large shoe size and have a high IQ.

f) Be a chain smoker and get lung cancer.

g) Dad is left handed and son is left handed.

4) A spinner with three equal spaces of red, blue, and green is spun one time. A single six-sided die is rolled once. What is the probability that you get blue and a number greater than 3?

5) Suppose you are dealt two cards, one after another from a standard deck of cards. What is the probability that both of your cards are:

a) spades?

b) the same suit?

c) kings?



6) Three cards are drawn from a standard deck without replacement. Find the probability that:

a) all are jacks.

b) all are clubs.

c) all are red cards.

7) In a carnival game, players are given three darts and throw them at a set of balloons on a wall. Suppose there are eight balloons on the wall. Five of the eight balloons have slips of paper in them that say 'Winner' while three of the eight balloons have slips of paper that are blank. Suppose you pop a balloon with each of your three darts. If all three balloons have 'Winner' slips, you win the grand prize. If all three balloons have blank slips, you win the consolation prize. What is the probability that:

a) you win the grand prize?

b) you win the consolation prize?

8) A classroom contains 12 males and 18 females. Two different students will be randomly selected to give speeches. What is the probability that the two students who give speeches are:

a) two females?

b) two males?

c) 1 male and 1 female (in either order)? (Hint: Use your answers from a) and b) along with some subtraction.)

9) If 18% of all Americans are underweight, find the probability that two randomly selected Americans will both be underweight.

10) A survey found that 68% of book buyers are 40 years old or older. If two book buyers are selected at random, what is the probability that both are 40 years old or older?

11) The Gallup Poll reported that 82% of Americans used a seat belt the last time they got into a car. If four people are selected at random, find the probability that they all used a seat belt the last time they got into a car.



12) Eighty-three percent of diners favor the practice of tipping to reward good service. If three restaurant customers are selected at random, what is the probability that all three are in favor of tipping?

13) Suppose that 25% of U.S. federal prisoners are not U.S. citizens.

a) Find the probability that a randomly selected federal prisoner is a U.S. citizen.

b) Find the probability that three randomly selected prisoners are all U.S. citizens.

14) At a local university, 70% of all incoming freshmen have computers. If three students are selected at random, what is the probability that:

a) none have computers?

b) all three have computers?

15) The U.S. Department of Justice states that 6% of all murders occur without weapons. If three murder cases are selected at random, what is the probability that all three occurred with the use of a weapon?

Review Exercises

16) Which of the following are random events?

i) You need to pick 2 people to be your partners in a group project so you select two of your friends.

ii) You make a rock skip across the surface of a lake 12 times.

iii) A baby elephant is born and it is a boy.

iv) You spin the big wheel on the TV game show "The Price is Right" and you win $1000.

17) In how many ways can two 12-graders be selected for speaking at graduation if there are 16 seniors that apply? One speaker will give a short introductory speech and one will give a longer speech that reflects upon the experiences of this particular senior class.



18) A family of 4 just won the lottery and goes to an auto dealership to purchase a new vehicle for each member of the family. The parents each decide that they want a car while their kids decide they would each like a truck. In how many ways can their purchase their 4 vehicles if the dealership has 17 cars and 23 trucks available?

19) The Strikers and the Kicks soccer teams are playing a best of five playoff series. The first team to win three games is the winner. Draw a tree diagram to show the different ways the series might play out.

2.3 Mutually Exclusive Outcomes

Learning Objectives

- Understand when two outcomes are mutually exclusive
- Understand the concepts of unions and intersections
- Be able to compute probabilities using Venn Diagrams and formulas

Sometimes there are situations in which two different outcomes cannot occur at the same time. For example, if you roll a single die one time and you wish to find the probability of getting an even number and a 3 on that one roll. These two outcomes cannot occur at the same time. When it is impossible for two outcomes to occur at the same time, we say the outcomes are **mutually exclusive** or **disjoint.** If outcomes 'A' and 'B' are mutually exclusive then it is impossible for outcome 'A' and 'B' to happen at the same time, or P(A and B)=0. However, if 'A' and 'B' are mutually exclusive then P(A or B)=P(A)+P(B). Using proper notation we have $P(A \cup B) = P(A) + P(B)$. *Remember, this is only true if the two outcomes are mutually exclusive.*

For Mutually Exclusive Events

$$P(A \cup B) = P(A) + P(B)$$

$$P(A \text{ or } B) = P(A) + P(B)$$

The $\cup$ can be read as the **union** of outcomes 'A' and 'B'. The probability for the union of two outcomes 'A' and 'B' can be thought of as the chance that either 'A' occurs, 'B' occurs, *or both 'A' and 'B'* occur.

Example 1

A single die is rolled one time. What is the probability of getting either an odd number or a 6?

Solution

The outcomes of getting an odd number and getting a 6 are mutually exclusive since they cannot occur at the same time. $P(Odd \cup 6) = P(Odd) + P(6) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \approx 0.67$

Of course, not every situation involves mutually exclusive events.

The diagrams above are **Venn Diagrams**. They are useful in showing us how different outcomes are related. Outcomes 'A' and 'B' are not mutually exclusive if they overlap. We say that there is an

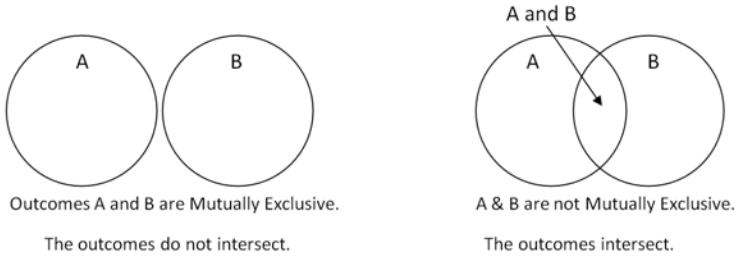


Figure 2.1

intersection of outcomes 'A' and 'B' if they overlap. For example, if we roll a single 6-sided die, the outcomes of getting an odd number and getting a number bigger than 3 intersect. They both include the number 5. The symbol for an intersection is $\cap$. A logical extension of the formula for the union of two outcomes that are not mutually exclusive is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. The next two examples illustrate this formula.

For Non-Mutually Exclusive Events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

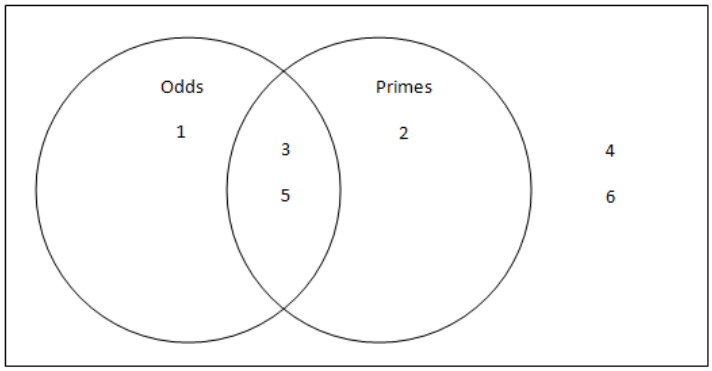
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



Example 2

A single 6-sided die is rolled. Suppose the outcomes we are interested in are getting an odd number and getting a prime number (2,3, or 5). Draw a Venn Diagram for this situation.

Solution



Notice that since the numbers 3 & 5 belong to both the odd numbers and the prime numbers, they are placed into the intersection of the 'Odds' circle and the 'Primes' circle. Notice also that the numbers 4 & 6 do not belong to either set and are placed outside both circles.

Example 3

A single 6-sided die is rolled. What is the probability of getting either an odd number or a prime number? Note that this is the same as asking for $P(Odd \cup Prime)$.

Solution

Using the figure from Example 1 we see that there are four values out of six that are either odd or prime. Therefore, P(Odd or Prime)$=\frac{4}{6} = \frac{2}{3} \approx 0.67$. If we use the formula, $P(Odd \cup Prime) = P(Odd) + P(Prime) - P(Odd \cap Prime)$. This gives $\frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6} = \frac{2}{3} \approx 0.67$.

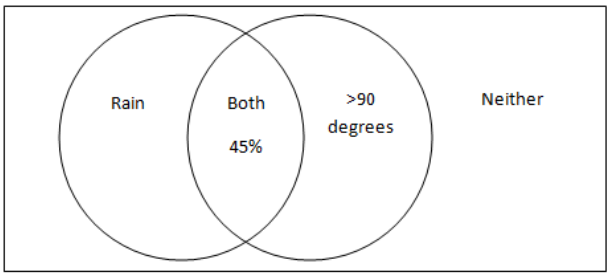
Example 4



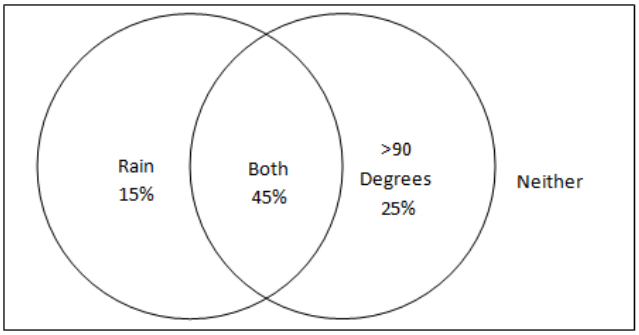
Suppose there is a 60% chance it will rain today and that there is a 70% chance that it will be over 90°F. Suppose also that there is a 45% chance that it will both rain and be above 90 degrees. What is the chance that it will neither rain nor be above 90 degrees? Solve using both a Venn Diagram and using a formula.

Solution

Start by drawing a Venn Diagram and noting that we have two circles, one for rain and one temperature. These two outcomes are not mutually exclusive because they can occur at the same time therefore the circles should overlap. We can quickly fill in the intersection of the two outcomes as 45%.



If we remember that there is a 60% chance of rain and we already have 45% filled in for the rain circle, the remainder of the rain circle must be 15% so it adds up to 60%. Likewise, the remainder of the >90° circle must be 25%.



We can now see that we have a total of 15%+45%+25%=85%. This means that the 'Neither' category must be 15% to give us a total of 100%.

Using the formula, $P(Rain \cup 90) = P(Rain) + P(90) - P(Rain \cap 90)$. Filling in we get $P(Rain \cup 90) = 60\% + 70\% - 45\% = 85\%$. 100%-85%=15%.

Example 5

Which pairs of outcomes are mutually exclusive?

a) You go to the pet store to buy a pet. Outcome A = You buy a pet that flies, Outcome B = You buy a pet that has no legs.

b) You order a pizza. Outcome A = Your pizza has pepperoni on it, Outcome B = Your pizza has mushrooms.

c) You select a football player to take a picture of for the yearbook. Outcome A = The player is a 4-year varsity starter, Outcome B = The player is 14 years old.

d) Radio stations have 4-letter station names such as KDWB. You decide to pick a radio station to listen to. Outcome A = The station's 4-letter name starts with a W, Outcome B = The station's 4-letter name contains three E's.

Solution

a) is mutually excluse. You cannot buy a pet that both flies and also has no legs.

b) is not mutually exclusive. You can order a pizza with both mushrooms and pepperoni.

c) is mutually exclusive. If a 4-year varsity starter is 14 years old, then they would have been a varsity starter when they were 10 years old. That simply does not happen.

d) is mutually exclusive. These both could happen if the radio station's call sign was WEEE.

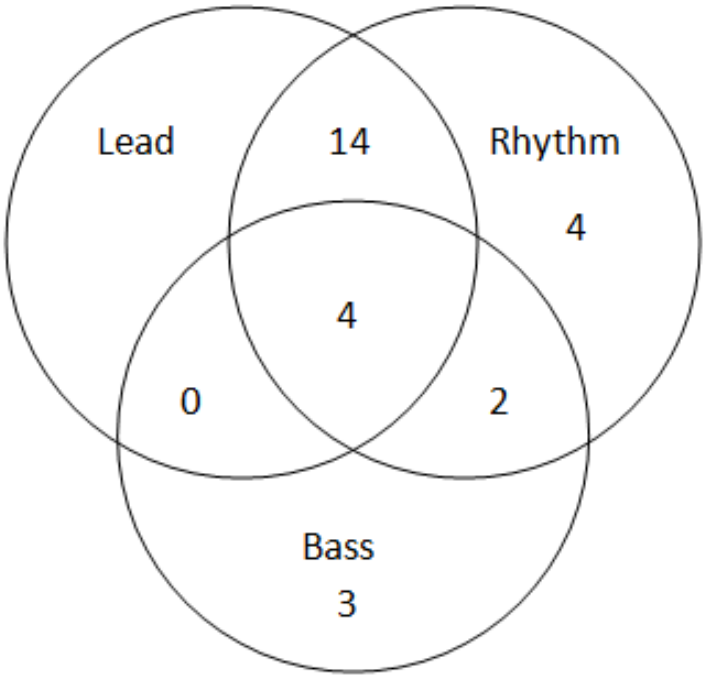
Example 6

The Rockin' Rollers performance company books shows for their musicians. Currently, they have 30 musicians who play either bass guitar, lead guitar, or rhythm guitar. Some of these musicians play more than one instrument. Suppose 4 musicians can play lead, rhythm, or bass guitar. Fourteen can play lead or rhythm but not bass, two can play bass or rhythm but not lead, 3 can play bass only, and 4 can play rhythm only. There are no musicians who play lead and bass only. Draw a Venn Diagram to determine how many musicians play lead only.



Solution

From the information given, we can fill in a Venn Diagram as shown below.



Notice that we have used up $14 + 4 + 4 + 2 + 3 = 27$ of the 30 musicians so that means there must be 3 musicians who play lead guitar only.

Problem Set 2.3

Exercises

1) What does it mean for two outcomes to be mutually exclusive?

2) Give an example of two outcomes from a situation in which the two outcomes are mutually exclusive.

3) Give an example of two outcomes from a situation that are not mutually exclusive.

4) Consider each event. Decide whether each pair of outcomes are mutually exclusive.

a) Roll a die: Get an even number and get a number less than 3.

b) Roll a die: Get a prime number (2, 3, or 5) and get a six.

c) Roll a die: Get a number greater than 3 and get a number less than 3.

d) Select a student: Get a student with blue eyes and get a student with blond hair.

e) Select a college student: Get a sophomore and get a student that is a math major.

f) Select a course: Get an Algebra course and get an English course.

g) Select a voter: Get a Republican and get a Democrat.

5) There are 200 male students at a particular school. Of these, 58 play football, 40 play basketball, and 8 play both.

a) Draw and label a Venn diagram for this situation.

b) How many play both sports.

c) How many play basketball but not football?

d) How many play football but not basketball?

e) How many do not play football or basketball?

6) A single card from a standard deck can have many descriptions. For example, the King of Spades could be described as a black card, a face card, a king, or a spade. Suppose we pull a single card out of a deck and we pay attention to the outcomes of getting a red card, getting a jack, and getting a spade.

a) Draw a Venn diagram to illustrate this situation paying attention to whether or not it is red, a jack, or a spade.

b) Shade the portion of your diagram with vertical lines that represents the intersection of getting a red card and getting a jack.

c) Shade the portion of the diagram with horizontal lines that represents the union of getting a jack or getting a spade.

7) An architectural firm is putting out bids to design two large governmental buildings. Suppose they believe they have 35% chance of getting the contract for the first building, an 80% chance of getting the contract for the second building and a 10% chance of getting neither job.

a) Draw a Venn Diagram for this situation and use your diagram to find the chance that they get both contracts.

b) Use a formula for this situation to find the chance that they get both contracts.

8) A student tells their teacher that they want to build a cabinet in woodshop. Students sometimes build this project with oak only, sometimes with cherry only, sometimes with both and sometimes with neither. There is a 40% chance the project will be built using oak, a 50% chance the project will be built using cherry, and a 30% chance that the project will be built using both types of wood. What is the chance that the student will not use either oak or cherry?



9) Consider a set of 15 pool balls. Balls numbered 1 through 8 are solid and balls 9 through 15 are striped. Suppose the balls are placed into a bag and one ball is randomly selected. Find the probability that:

a) you selected either a solid ball or a ball numbered greater than 12?

b) you selected an even numbered ball or a solid ball?

c) you selected a solid ball or a striped ball?

d) you selected a ball that was striped and even?

10) Suppose you again have a standard set of 15 pool balls. This time, you pull two pool balls out of the bag, replacing the first ball before you select the second ball. What is the probability that:

a) your two pool balls are both solid?

b) you pick exactly the same ball twice?

c) your first ball was solid and your second was odd?

11) At a particular school, there are 20 teachers. Three of them teach math, 5 teach science, and 3 teach computer science. It turns out that there is one teacher who teaches all three classes and one teacher who teaches both science and computer science. Draw a Venn diagram to illustrate the situation and determine how many of the 20 teachers teach courses other than math, science, or computer science. Hint: You will need 3 circles to build this diagram.

Review Exercises

12) Two cards are selected from a standard deck of 52 cards, one after the other without replacement. What is the probability that the two cards are both face cards?

13) Suppose 90% of all Americans have attended a religious ceremony at least one time in the past year. What is the probability that 4 randomly selected Americans will all have attended at least one religious ceremony in the past year?



14) A single 6-sided die is rolled once and a single card is drawn from a standard deck of 52 cards. What is the probability that the die shows a result greater than 3 and the card is a heart?

15) A young girl has a box of 8 color crayons but has decided they need only 3 colors to make a picture for her grandfather. In how many ways can the child select the three crayons?

16) In how many ways can a committee of 4 people be selected if there must be at least 1 man and 1 women on the committee and there are 6 men and 7 women from which to pick?

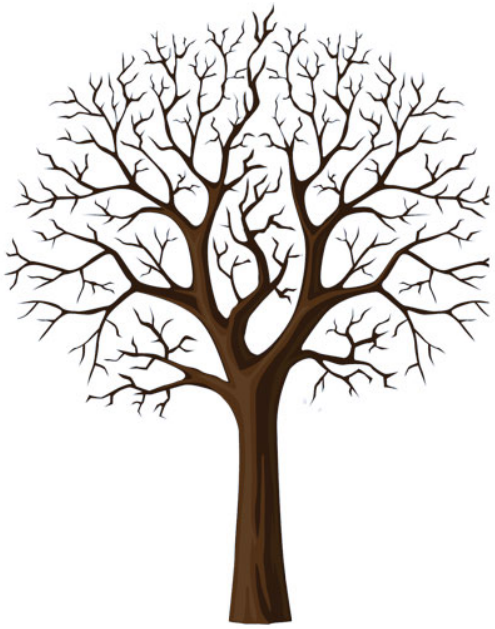
2.4 Tree Diagrams and Probability Models

Learning Objectives

- Understand how to build and properly notate a tree diagram
- Understand how to calculate probabilities using a tree diagram
- Understand how to verify if a tree diagram is correct

- Be able to build a probability model by using a tree diagram

As we advance through probability, it becomes very apparent that we need to be quite organized with our problems as they become more complex. In this section we will use tree diagrams to help us calculate probabilities for given situations. **Tree diagrams** are a visual aid that can help us break down a situation and calculate probabilities. There are two key principles that we must observe for all tree diagrams. First of all, to find the total probability for any given branch on a tree, multiply the individual probabilities along that branch. Secondly, the sum of the probabilities from the ends of each branch must total to 1. We will examine several examples of probabilities using tree diagrams in order to solidify our understanding of this concept.

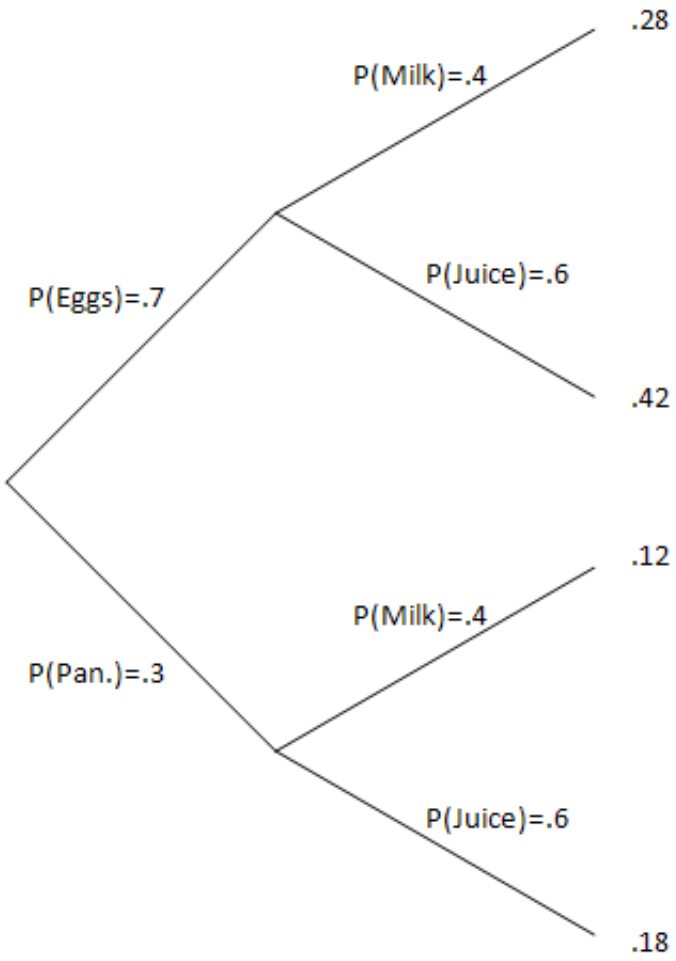


Example 1

At a restaurant, there are two breakfast platters that are served, one featuring pancakes and one featuring eggs. There are also two choices for drinks, milk or juice. Thirty percent of customers choose the pancake platter while 70 percent choose the egg platter. Forty percent of customers choose milk while 60 percent choose juice. Assume the drink choice is independent of platter choice. Build a probability model for this situation by using a tree diagram.

Solution

Step one is to build the tree diagram as shown below. Be sure to label each branch with what it represents and the associated probability. Step two is to calculate the probabilities at the end of each branch. To do this, we multiply the probabilities along each branch. For example, the top branch's value of 0.28 was found by multiplying 0.7 by 0.4.



Our probability model, shown below, summarizes the data in the tree diagram. There are two critical ideas to pay attention to here. First of all, the probability at the end of each branch is the product of the probabilities on that branch. Secondly, notice that the sum of the four probabilities at the ends of the branches add up to 1. The table below that summarizes the results from our tree diagram and is called a **probability model**. Notice that the probabilities in the table sum to 1.

Table 2.1:

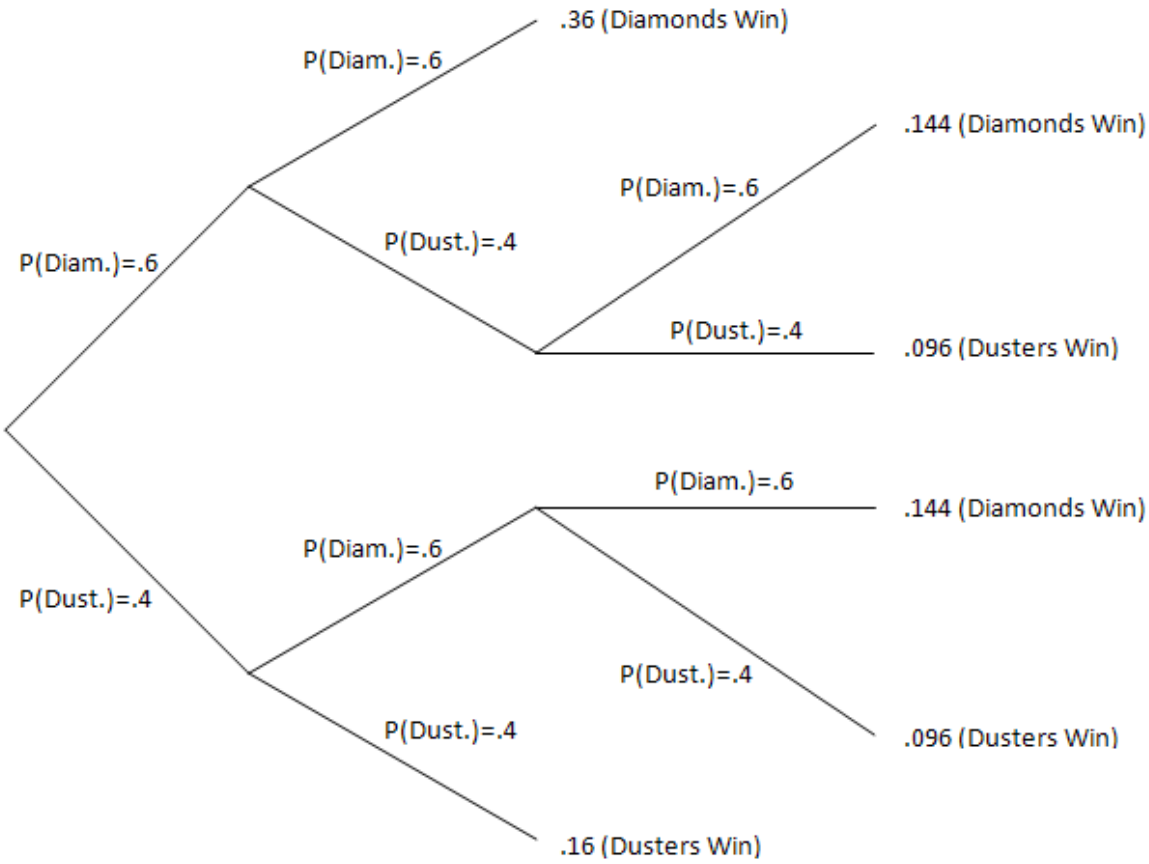
Order	Eggs & Milk	Eggs & Juice	Pancakes & Milk	Pancakes & Juice
Probability	0.28	0.42	0.12	0.18

Example 2

The Diamonds and the Dusters baseball teams are going to play a best-of-three playoff series. The first team to win two games is the winner of the series. Based on previous performance this season, the Diamonds have a 60% chance to win any game they play against the Dusters. Build a tree diagram and then build a probability model to determine the probability of each team winning the series.

Solution

Notice that in the tree diagram below, not all branches go three games. There are two sets of branches that only go two games. A third game does not need to be played in all situations.



We can now make a probability model for this situation by adding up the probabilities for the Diamonds winning in two or three games and the Dusters winning in two or three games. This is shown in the probability table below. The probability of the Diamonds winning the series is .36+.144+.144=.648. The probability of the Dusters winning the series is .096+.096+.16=.352. In our solution, we were careful to make sure each branch was labeled and included a probability. Once again, we multiplied the values along each branch to get the probability at the end of the branch. Notice again that the total of all the probabilities in the probability model adds to 1.

Table 2.2:

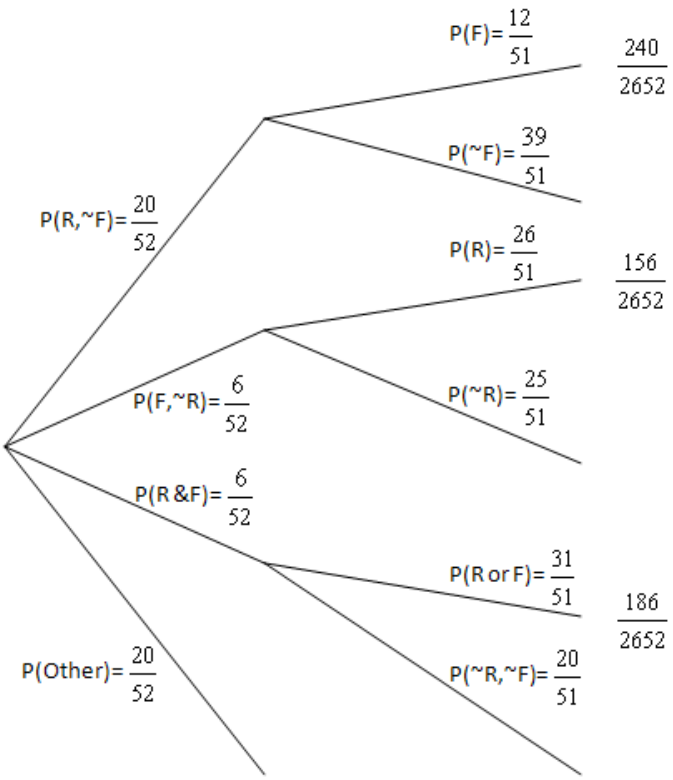
Outcome	Diamonds Win in 2 games	Dusters Win in 2 games	Diamonds win in 3 games	Dusters win in 3 games
Probability	0.36	0.16	0.288	0.192

Example 3

You are dealt two cards from a standard deck of 52 cards. What is the probability that the two cards can be classified as a red card and a face card (in either order)?

Solution

Begin by considering the first card. We are concerned primarily with getting a face card and then a red card or a red card and then a face card. Let's start by thinking about what we would need on our first card to have a chance to reach our goal. The first card could be a red card, a face card, or both a red and a face card. Any other card would result in us not reaching our goal. As a result our tree diagram will have four initial branches as shown below; red (not face), face (not red), red & face, or other.



Once our first set of branches are complete, we look at the second stage. We will examine the red, not face branch in detail (top branch). There are 20 red cards out of our 52 card deck that are not face cards. Once we have that card, we now want to know the chances of getting a face card (assuming we just drew a red card). There are still 12 face cards in the deck but there are now only 51 cards remaining in the deck. We multiply this branch out to get $\frac{20}{52} \times \frac{12}{51} = \frac{240}{2652}$. There are two other branches we work in a similar fashion as they are the only other branches that help us achieve our goal. Adding these gives us $\frac{240}{2652} + \frac{156}{2652} + \frac{186}{2652} = \frac{582}{2652} \approx 0.22$. There is about a 22% chance of getting a red card and a face card.

The key to success when working with tree diagrams and probability models is to work in a neat and organized fashion. Many errors are often due to sloppy work. In addition, another useful suggestion is to make your tree diagrams large enough so that you have plenty of room to work.

Problem Set 2.4

Exercises

1) Fourteen red marbles and sixteen green marbles are in a bag. Two marbles are picked out one at a time and replaced after they are picked. Build a tree diagram and probability model to show the different combinations of marbles that could be pulled out of the bag.

2) A bag contains a standard set of pool balls. Two balls are pulled out, one after another, and not replaced. What is the probability that the two balls are a solid and a striped ball in either order? (Recall that there are 8 solid pool balls and 7 striped pool balls.)

3) A bag contains a $100 bill and two $20 bills. A second bag contains 1 gold marble and 2 silver marbles. You get to pick one bill out of the first bag. After this, you pick a marble out of the second bag. If you get the gold marble, you get to triple the amount of money you pulled from the first bag. If you get a silver marble, you get to double the amount of money you picked from the first bag. Build a probability model for all the different amounts of money that you might win.



4) A basketball player is practicing shooting free throws. Suppose she makes 75% of her free throw attempts. Make a tree diagram and probability model for what might happen if she decides to shoot three free throws. In other words, what is the probability that she makes zero shots, one shot, two shots, or all three shots.

5) A coin is flipped and then two dice are rolled. Build a probability model that shows how likely it is to get heads followed by doubles, heads and a non-doubles, tails and doubles, and tails and non-doubles.

6) A spinner with four evenly-spaced wedges of red, blue, green, and orange on it is spun and a coin is flipped.

a) How many different outcomes are possible?

b) Build a probability model that shows the probabilities for each outcome.

7) A baseball player is a .400 hitter. This means that he gets a hit (single, double, triple, or home run) 40% of the time he has an at-bat. Use a tree diagram to build a probability model that shows the probability of the player having 0, 1, 2, or 3 hits if he has 3 at-bats in one game.

8) In some sports, the home team wins a higher percentage of games played. Suppose the Dunkers and the Hoopsters are playing a best-of-three game series against each other. When the Dunkers are home, they

have a 60% chance of winning a game against the Hoopsters. When the Hoopsters are home, they have a 55% chance of winning a game against the Dunkers. The Dunkers will be the home team in games 1 and 3 while the Hoopsters will be the home team in game 2. Use a tree diagram to build a probability model for this situation. The model should show the chances that the Dunkers win in 2 games or in 3 games and the chances that the Hoopsters win in 2 games or in 3 games.

9) A patient is scheduled to have two surgeries. The results of each surgery are independent of each other. Suppose the first surgery has a 90% success rate and the second surgery has an 85% success rate. Build a probability model by using a tree diagram that shows all the different results that might occur.



10) A bag contains ten red cubes numbered 1 through 10 and five blue cubes numbered 1 through 5. You pull two cubes out of the bag without replacement. What is the probability that the two cubes will be an odd cube and a red cube (in either order)?

Review Exercises

11) Suppose events 'A' and 'B' are mutually exclusive and that P(A)=0.35 and P(B)=0.14. What is $P(A \cup B)$?

12) How many unique three-letter 'words' can be formed by selecting three letters from the alphabet if no letter may be repeated?

13) How many unique three-letter 'words' can be formed by selecting three letters from the alphabet if letters may be repeated?

14) 20% of all households in the Twin Cities get the Star Tribune newspaper delivered to their home while only 15% get the Pioneer Press delivered to their home. If 70% of homes do not get either newspaper delivered, what percent of homes get both newspapers delivered?

2.5 Conditional Probabilities and 2-Way Tables

Learning Objectives

- Understand how to calculate conditional probabilities
- Understand how to calculate probabilities using a contingency or 2-way table

It is quite easy to calculate simple probabilities. What is the chance of rolling a 4 with a single die? What is the chance of being dealt a queen from a deck of cards? We are now going to focus on conditional probabilities. A **conditional probability** is a probability in which a certain prerequisite condition has already been met.

We can start by thinking about cards being dealt from a standard deck of 52 cards. As each card is dealt, what remains in the deck changes. A gambler in a casino will pay close attention to cards played. If many face cards have already been dealt, the observant gambler will understand that the next card has a higher chance of not being a face card. Suppose we want to know the probability that our next card will be a face card given that the first card was the 7 of diamonds. The formal notation for this is P(Face|7♦). This is read as "The probability of a face card given that we already have been dealt the 7 of diamonds.". Often times the math for these situations is very logical. In our case, we have simply reduced the deck by one card and there are still 12 face cards in the deck. Therefore P(Face|7♦)=$\frac{12}{51} = \frac{4}{17} \approx 0.24$.

Example 1

Two cards are dealt from a standard deck of 52 cards. Find each conditional probability.

a) P(2nd red|1st 2♣)

b) P(2nd red|1st 2♦)

c) P(2nd club|1st red)

Solution

a) There are still 26 red cards out of the remaining 51 cards after the two of clubs is dealt. $\frac{26}{51} \approx 0.51$.

b) There are only 25 red cards out of the remaining 51 cards after the two of diamonds is dealt. $\frac{25}{51} \approx 0.49$.

c) There are still 13 clubs out of the remaining 51 cards after a red card is dealt. $\frac{13}{51} \approx 0.25$.

Example 2

In a common poker game, 5 cards are dealt to a player. The best possible hand is called a royal flush. This occurs if a player gets the ten, jack, queen, king, and ace all of the same suit. What is the chance of being dealt a royal flush? Leave your answer as a fraction.

Solution



We will solve this by looking at one card at a time. What is the chance that the first card might be part of a royal flush? Before any cards are dealt, there are four 10's, four jack's, four queen's, four kings, and four aces available. Twenty of the 52 cards can help you on your way to a royal flush.

Once you receive this card, what are the chances that the second card will also help on your way to the royal flush? We might answer this by simply imagining us getting a useful card on the first card. Suppose our first card was the jack of spades. There are only 4 other cards of the remaining 51 cards that will help now, the 10, queen, king, and ace of spades. Suppose we get one of those cards, perhaps the king of spades.

There are now only 3 cards of the remaining 50 that can help us complete our royal flush. Suppose our third card was the queen of spades. Only 2 of the remaining 49 cards will help our quest for our fourth card. Likewise, there is only one card of the last 48 that can help us on card number five. Putting this all together, we have $\frac{20}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48} = \frac{480}{311{,}875{,}200} = \frac{1}{649{,}740}$. Putting this in perspective, if you dealt 1000 poker hands every single day, it would take nearly two years to deal 649,740 hands, of which we would only expect about one to be a royal flush. Good Luck!

Another way we can look at conditional probabilities is through the use of **two-way tables** or **contingency tables**. These are often referred to as two-way tables because there are two distinct pieces of information gathered in these tables. For example, we may record how many siblings you have and in how many activities you participate in school. Two-way tables can be filled in either using counts or probabilities.

We will start by answering simple questions such as "What is the probability that a student participates in exactly 2 activities?". After we understand how to work with these tables, we will begin asking more complex questions such as "What is the chance a student participates in 3 activities given that they have 1 sibling?". Let's begin with an easy example to help us understand how to read these tables.

Example 3

Suppose we survey all the students at school and ask them how they get to school and also what grade they are in. The chart below gives the results. Suppose we randomly select one student.

	Bus	Walk	Car	Other
9th or 10th grade	106	30	70	4
11th or 12th grade	41	58	184	7

a) Give all the row and column totals.

b) What is the probability that the student walked to school?

c) What is the probability that the student was a 9th or 10th grader?

d) What is the probability that a student either rode the bus or is in 11th or 12th grade?

Solution

a)

	Bus	Walk	Car	Other	Total
9th or 10th grade	106	30	70	4	210
11th or 12th grade	41	58	184	7	290
Total	147	88	254	11	500

b) There were 88 walkers out of 500 total students or $\frac{88}{500} = \frac{22}{125} \approx 0.18$.

c) There were 210 9th or 10th graders out of 500 total students or $\frac{210}{500} = \frac{21}{50} = 0.42$.

d) There are 147 kids who rode the bus and there are 290 kids who are 11th or 12th graders. However, notice that these two categories intersect and we must be careful not to count the 41 kids who are in both categories twice. We will take the 290 11th or 12th graders and just add the 106 bus riders who are not 11th and 12th graders for a total of 396 students. The probability of selecting an 11th or 12th grader or a bus rider is $\frac{396}{500} = \frac{99}{125} \approx 0.79$.

In the example above, note that the total across the bottom, 147+88+254+11, and the total for the last column, 210+290, both add up to 500. This is true of all 2-way tables. Now that we have the basic ideas down in a contingency table, let's move to a couple of more challenging questions.

Example 4

Consider the completed chart in the solution of part a) of Example 3.

a) What is the probability that a student is in 11th or 12th grade *given that* they rode in a car to school?

b) What is P(Walk|9th or 10th grade)?

Solution

a) The trick to dealing with conditional probabilities in two-way tables is to make sure that you only use what you are given. We are given that they rode in a car to school. *We will only look at the Car column.* We first note that there were a total of 254 kids who rode in a car to school. We then see that 184 of these kids were 11th and 12th graders. This gives us $\frac{184}{254} = \frac{92}{127} \approx .72$.

b) We want the probability that a student walked to school given that they were in 9th or 10th. *We will only look only at the 9th and 10th grade row.* There are 210 students who are 9th and 10th graders. Of these, only 30 walked to school. This gives us $\frac{30}{210} = \frac{1}{7} \approx .14$.

Example 5

The manager of an ice cream shop is curious as to which customers are buying certain flavors of ice cream. He decides to track whether the customer is an adult or a child and whether they order vanilla ice cream or chocolate ice cream. He finds that of his 224 customers in one week that 146 ordered chocolate. He also finds that 52 of his 93 adult customers ordered vanilla. Build a contingency table that tracks the type of customer and type of ice cream.

Solution

Start by filling in the values we are given and then work from there. The table below shows what we are given in the initial problem.

	Adult	**Child**	**Total**
Vanilla	52		
Chocolate			**146**
Total	**93**		**224**

Our next step is to fill in the Adult/Chocolate space with 41, the Child/Total box with 131, and the Total/Vanilla box with 78 by using subtraction. It is now easy to fill in the remaining

boxes. For example, we can quickly determine that the Child/Chocolate box must be 146 - 41 = 105. You can verify that this table is correct by checking each row and column total.

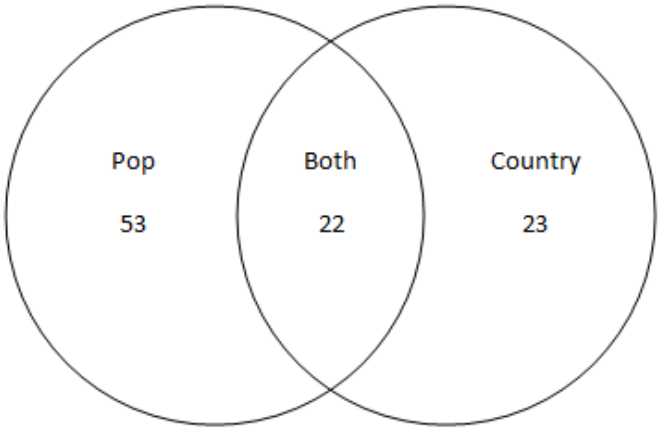
	Adult	Child	Total
Vanilla	52	26	**78**
Chocolate	41	105	**146**
Total	**93**	**131**	**224**

Example 6

A survey asked students which types of music they listen to? Out of 200 students, 75 indicated pop music and 45 indicated country music with 22 of these students indicating they listened to both. Use a Venn Diagram to find the probability that a randomly selected student listens to pop music given that they listen country music.

Solution

Consider our Venn Diagram below. First fill in the both section with 22 students. We can now logically deduce how many students are left to fill up the Pop circle and the Country circle. Since we are given that they listen to country music we may only use the information that is in the Country circle. There are only 45 students that landed in this circle. Of the 45 students who listen to country music, 22 of them also listen to pop music or $\frac{22}{45} \approx .49$.



Problem Set 2.5

Exercises

1) Figure 2.2 shows the counts of earned degrees for several colleges on the East Coast. The level of degree and the gender of the degree recipient were tracked. Row & Column totals are included.

	Bachelor's	Master's	Professional	Doctorate	Total
Female	542	128	26	18	**714**
Male	438	165	38	20	**661**
Total	**980**	**293**	**64**	**38**	**1375**

Figure 2.2

a) What is the probability that a randomly selected degree recipient is a female?

b) What is the probability that a randomly chosen degree recipient is a man?

c) What is the probability that a randomly selected degree recipient is a woman, given that they received a Master's Degree?

d) For a randomly selected degree recipient, what is P(Bachelor's Degree|Male)?

2) In poker, 5 cards are dealt to a player. One of the stronger poker hands is a flush. This means that all 5 cards are of the same suit, for example, all hearts. What is the probability of being dealt a flush?

3) The table below shows the probability breakdown of ages and genders for the typical American college student. Each value in the table is given as a probability. For example, there is a 12% chance that a randomly selected college student will be a male between 25 and 34 years old.

Table 2.3:

	14-17	18-24	25-34	>34
Male	.01	.30	.12	.04
Female	.01	.30	.13	.09

a) What is the probability that a randomly selected American college student is female?

b) What is the probability that a randomly selected American college student is female given that the student is more than 34 years old?

c) What is the probability that a randomly selected college student is either a female or more than 34 years old?

4) Suppose that 40% of adults like eating bananas while 60% like eating apples. Suppose also that 32% of adults like eating both. What is the conditional probability that a randomly selected adult likes apples given that they like bananas? Use a Venn Diagram to answer this question.

5) Another good poker hand is called a straight. This means that your five cards will be numerically in order such as an 8, 9, 10, jack, and queen. The cards do not need to match suit in a straight. Suppose you receive the first four cards of a five card poker hand. You have 5♥, 7♦, 8♣, and 9♦. What is the probability that the next card will give you a straight?



6) Suppose you receive the first four cards of a five card poker hand. You have 3♥, 4♦, 5♣, and 6♦. What is the probability that your next card will give you a straight?

7) A statistics class has 18 juniors and 10 seniors in it. 6 of the seniors are females and 12 of the juniors are males. Build a contingency table to find the probability that a randomly selected student is:

a) a junior or a female?

b) a senior or a female?

c) a junior or a senior?

d) a female given that the student was a senior?

8) At a used-book sale, there are 120 children's books and 80 adult books available. 50 of the adult books are nonfiction while 40 of the children's books are nonfiction. All other books are fiction. Build a contingency table to find the probability that a randomly selected book is:

a) fiction.

b) not a children's nonfiction

c) an adult book or children's nonfiction.

d) a children's book given that it was nonfiction.

9) Cable channels 6, 8, & 10 show quiz shows, comedies, & dramas. The table below shows the distributions of these shows.

Table 2.4:

	Channel 6	Channel 8	Channel 10
Quiz Show	4	2	1
Comedy	3	3	8
Drama	4	5	1

If a show is selected at random, find the probability that the show is:

a) a quiz show or shown on Channel 8.

b) a drama or a comedy.

c) a comedy given that it is shown on Channel 8.

d) shown on Channel 6 given that it is a drama.

10) Animals on the endangered species list are given in the table below by type of animal and whether it is domestic or foreign to the United States.

Table 2.5:

	Mammals	Birds	Reptiles	Amphibians
United States	63	78	14	10
Foreign	251	175	64	8

An endangered animal is selected at random. What is the probability that it is:

a) a bird found in the United States?

b) foreign or a mammal?

c) a bird given that it is found in the United States?

d) a bird given that it is foreign?



11) Suppose a standard set of pool balls (1-8 are solid and 9-15 are striped) are in a bag. Two balls are picked out of the bag without replacement.

a) Find the probability that second ball is striped given that the first ball was the 10 ball.

b) Find P(2nd striped|1st solid).

12) Suppose you receive your first three cards of a five card poker hand. You have 5♦, 6♦, 7♥. What is the probability that your next two cards will result with you having a straight?

Review Exercises

13) Suppose that 25% of all 9th graders have an unweighted GPA after their 9th grade year of 3.5 or higher. Also suppose that 60% of all 9th graders are involved in a sport at some time during their 9th grade year. Assume that a student's GPA and whether or not they are in a sport are independent of one another. Draw and label a tree diagram for this situation and build a probability model that summarizes the different probabilities possible for 9th grade students in regards to GPA and a sport.

14) A special deck of cards contains only the face cards and aces from a standard deck of cards.

a) If one card is dealt, what is the probability that the card is an ace?

b) If one card is dealt, what is the probability that the card is a black ace?

c) If two cards are dealt, what is the probability that both cards are face cards?

15) Suppose for a moment that all months have exactly 30 days and the chance of you being born in any particular month is $\frac{1}{12}$. What is the probability that neither of two randomly selected people will have been born in the same month as you?

16) The standard California license plates made in 2011 or later must begin with a digit between 6 or 9 followed by a letter anywhere from T to Z. They then have any two letters followed by any three digits. How many of these license plates are possible?

2.6 Chapter 2 Review

Probability in its most basic form is a simple question of how likely it is for an outcome to occur. We are always looking to divide the number of favorable outcomes by the total number of outcomes. We cannot

predict a specific outcome for a random event but the Law of Large Numbers allows us to make long term predictions of chance behavior. There are numerous rules dictating how we must calculate probabilities. These rules often deal with whether events are independent, outcomes are mutually exclusive, or whether replacement is used. We also must deal with conditional probabilities for situations in which a particular outcome is assumed to have occurred. To help us organize each situation, we can often utilize Venn Diagrams, tree diagrams, and contingency or 2-way tables. Having a clear understanding of situations and being organized when dealing with probability is critical when performing probability calculations.

Chapter 2 Review Exercises

1) Suppose that 57 of 110 students at a school are underclassmen (freshmen or sophomores) while the rest of the students are upperclassmen (juniors or seniors). Suppose three students are selected at random.

a) What is the probability that all three of the students are underclassmen?

b) What is the probability that all three students are upperclassmen?

c) What is the probability that there is at least one underclassman and at least one upperclassman in the group of three students?

2) Two 6-sided dice are rolled, one after the other. Find each probability.

a) P(total of 10 or more)

b) P(doubles)

c) P(total is even or less than 6)

d) P(an odd product)

e) P(first die is greater than second die)

f) P(a 6 or a three is showing on at least one die)

g) P(an odd total or a 2 is showing)

3) A pet store surveys his customers during the day and finds that 15 customers own dogs and 9 own cats. Included in these were 4 customers who owned both.

a) Draw a Venn Diagram for this situation

b) How many total customers were surveyed?

c) Suppose one of these customers was selected at random. What is P(owned a dog)?

d) Suppose one of these customers was selected at random. What is P(own a dog|own a cat)?

e) Suppose one of these customers was selected at random. What is P(own a cat|own a dog)?

3) Suppose that 40% of all adults in a certain town are females and that 60% are males. In addition, 60% of the females hold full-time jobs while 80% of the males hold full-time jobs.

a) Draw and label a tree diagram to represent this situation.

b) What is the chance that a randomly selected person holds a full-time job?

4) For Halloween at my house, kids spin a spinner that has three equally marked spaces labeled 1, 2, and 3. The number they spin is the number of pieces of candy they get. In my bag, I start with 20 chocolate bars and 30 sugar bombs - all with identical packaging. Trick-or-Treaters pick randomly out of my bag after they spin. I only restock my candy bag after each child finishes picking all their candy.



a) What is the chance that a trick-or-treater gets to pick three pieces of candy?

b) Suppose a trick-or-treater spins a three. What is the chance that they pick three sugar bombs?

c) Suppose a trick-or-treater spins a three. What is the chance that they pick three chocolate bars?

d) What is the chance that the trick-or-treater gets only one chocolate bar and nothing else?

e) What is the chance that the trick-or-treater gets exactly one chocolate bar and one sugar bomb?

5) Suppose the table below gives a breakdown of the ages and genders of the teachers at your school.

Table 2.6:

	$\leq$29	30-39	40-49	$\geq$50
Male	5	6	18	7
Female	7	7	13	4

Find the probability that a randomly selected teacher is:

a) a male.

b) 39 years old or younger.

c) either a male or at least 50 years old.

d) from 30 to 39 years old given that they are a female.

e) a female given that they are at least 40 years old.

6) Consider the 2-way table shown below that shows the number of different types of automobiles produced by major manufacturers.

Table 2.7:

	GM	Ford	Chrysler	Toyota
Cars	14	11	12	7
Trucks	8	9	5	6
Vans	2	3	5	3

What is the probability that a randomly selected vehicle is:

a) a Ford?

b) a truck?

c) a van or a Toyota?

d) a car given that the vehicle is built by GM?

e) a Ford given that the vehicle is a truck?

7) Two cards are dealt from a standard 52 card deck without replacement. What is the probability that the two cards are both face cards?

8) A baseball player has a batting average of .250 which means that he averages one hit for every four times he comes to the plate. What is the probability that this player will end up with exactly 2 hits if he comes to the plate 3 times in a single game?

9) Two bags have an assortment of marbles in them. The first bag contains 11 black, 12 white, and 7 gold marbles. The second bag contains 9 black and 11 white marbles. One marble is randomly selected out of each bag.

a) Draw a tree diagram to represent this situation.

b) What is the probability that the two marbles are both black?

c) What is the probability that the two marbles are the same color?



10) A special deck of cards contains only the eight red cards that are face cards or aces. Two cards are dealt off the top of the deck.

a) What is the probability that the two cards you end up with are both kings?

b) What is the probability that the two cards are of different value?

c)What is the probability that the two cards have the same value (two kings, two queens, etc...)?

d) What is the probability that the two cards are the same suit?

11) A bag contains ten red cubes numbered 1 through 10 and five green cubes numbered 1 through 5. Two cubes are pulled from the bag at random. What is the probability that the two cubes are:

a) both red?

b) both odd?

c) the same color?

d) the same value?

12) For a carnival game, a bag contains one $100 bills and nine $20 bills. You roll a single 6-sided die one time. If you roll a one or two you get to pull one bill out of the bag. If you roll a three, four, five, or six, you get to pull two bills out of the bag.

a) Draw a tree diagram for this situation.

b) Build a probability model for this situation.

b) What is the probability that you win exactly $120?

13) A burglar alarm system has three separate detection mechanisms it uses to detect an intruder. Suppose a skilled burglar has an 30% chance to get around the first part of the detection system, a 60% chance of getting around the second part of the system, and a 55% chance of getting around the third part of the system. Assume each part of the detection system is independent of the other parts of the system.



a) What is the chance that the system does not detect the burglar?

b) Based upon your answer to part a), what must be the chance that the system does detect the burglar?

c) What is the chance that the burglar can get around exactly two of the three detection systems

14) On a basketball team, players can play at least one of three positions; guard, forward, or center. Suppose that 30 girls try out for the basketball team. During tryouts 13 girls indicate they can play guard only, 3 state they can play center only, 6 state they can play center or forward and the rest state they can play forward only. A player is selected at random.

a) Draw a Venn diagram for this situation.

b) What is the probability that the randomly selected player says they can play forward?

c) Given that the player indicates they can play forward, what is the probability they can also play center?

15) A girl is deciding what jewelry to wear as she gets ready for school. She has 5 bracelets, 6 rings, and 8 necklaces from which to choose.

a) In how many ways can she choose exactly one of each item to wear from the 19 available items?

b) If she decides to randomly select three pieces of jewelry, what is the probability that all three of the items she picks are exacty the same type of jewelry?

c) What is the probability that she picks exactly one bracelet, one ring, and one necklace if she randomly selects three pieces of jewelry?

16) Your statistics teacher needs to select 3 students to help demonstrate an activity. Your class has 12 sophomores, 19 juniors, and 5 seniors in it. Your teacher makes a random selection of three students.

a) In how many ways can your teacher select three students from this class of 36 students?

b) What is the probability that all three students will be juniors?

c) What is the probability that exactly one student from each grade will be selected?

17) All football plays that an offense can run can be classified as a pass, run, or a kick. No play can ever be put into two categories.

a) If an offense completes two plays, will these two plays be independent of each other? Why or why not?

b) If the offense runs one play, are the possible outcomes (pass, run, or kick) mutually exclusive? Why or why not?

Image References

Coins http://coinauctionshelp.com

Pool Balls http://plutonius.aibrean.com

Scattegories Die http://ehow.com

October Calendar http://printablecalendars.resources2u.com

Roulette Wheel http://www.partypokersupplies.co.uk

Kids at Board http://teachers.greenville.k12.sc.us

Two Striped Pool Balls http://demo.physics.uiuc.edu

Ace and King of Spades http://www.123rf.com

Seat Belt http://sawmengzhi.blogspot.com

Graduation http://www.prlog.org

Aerosmith http://www.obit-mag.com

Cabinet http://www.renovation-headquarters.com

Cathedral in St. Paul, MN http://www.scenicreflections.com

Tree http://www.onenewsnow.com

$100 Bill http://onlinecurrencytradingfxcm.blogspot.com

Royal Flush http://www.artpoker.net

Straight http://www.findabet.co.uk

Turtle http://www.maine.gov

Trick or Treaters http://www.myremoteradio.com

Bag of Marbles http://www.worldwiseimports.com

Burglar http://www.emovingstorage.com

Chapter 3

Expected Values & Simulation

3.1 Probability Models & Expected Value



(Note- Three Separate Videos for the Section)

Learning Objectives

- Be able to construct a probability model (expected value table) given all possible outcomes and the associated probabilities
- Be able to calculate the expected value for a situation given a probability model
- Be able to calculate missing values in a probability model given information about the expected value in a situation.

Suppose you walk into a casino. You will see all sorts of games varying from blackjack and poker to slot machines. It would not take long for you to notice that there are some players who are winning some money, sometimes a substantial amount. You might wonder how the casino makes money when they are clearly giving some money away.



Casinos have a clear understanding of expected value. The **expected value** for a situation can be thought of as the average result over the long run. In other words, it can be thought of the expected winnings or average payout for a game of chance. Consider the thinking of the owner of a casino. While there are some people who win a little, and occasionally a few people who win a lot, most people end up losing some money at the casino. The casino actually expects some people to occasionally win big. In fact it makes for great advertisement! As long as the mathematics show that the expected value is in the casino's favor, the casino will continue to make money in the long run. In this section, we will focus on how to calculate the expected value.

As mentioned above, the expected value can be thought of as the average result over the long run. Recall that to find the average value of a series of numbers, we simply add up the numbers and divide by how ever many numbers there are. For example, the average of 3, 4, 5, and 6 is 4.5 because $\frac{3+4+5+6}{4} = 4.5$. You will notice that the average value of 4.5 is not one of the numbers in the original set of numbers. This is often also true with expected values. *The expected value for a situation does not have to be one of the possible values.*

Use the concept of averages to find the expected value for the example below.

Example 1

A game is played in which a coin is flipped one time. If the coin lands on heads, the player wins \$5. If the coin lands on tails, the player wins \$10. What is the expected value for a player who plays this game one time?

Solution

The expected value is \$7.50. This is strange because it is actually impossible for a player to win \$7.50. They could only win either \$5 or \$10 but the average win will be \$7.50. One way to see this is to actually play the game two times. If the flips come out matching their theoretical probabilities, one of the flips will be heads for \$5 and the other will be tails for \$10. The player will have won \$15 in two games so the average win or expected value would be $\frac{\$5+\$10}{2} = \frac{\$15}{2} = \$7.50$.

This method works quite well in simple situations, but it gets more cumbersome as the situations get more complex. Consider the example below.

Example 2

Student council is raising money to support a program called "Shoes for the Homeless". A booth was set up in the lunchroom at which students could pledge a donation of $1, $5, or $10 for money towards a large shoe purchase. 125 students pledged money for this fundraiser. Eighty students pledged $1, 25 students pledged $5, and 20 students pledged $10.

a) Build a probability model for this situation.

b) What was the average donation per student?

Solution

a) Remember that a probability model needs probabilities, not just counts. For $1 pledges we have $\frac{80}{125} = 0.64$, for $5 pledges we have $\frac{25}{125} = 0.2$, and for $10 pledges we have $\frac{20}{125} = 0.16$. Notice that 0.64+0.2+0.16=1 or 100%.

Value	$1	$5	$10
Prob.	0.64	0.2	0.16

b) There were 80 students who pledged $1 each for a total of $80, there were 25 students who pledged $5 each for a total of $125, and there were 20 students who pledged $10 each for a total of $200. All the pledges added together give us $\$80 + \$125 + \$200 = \$405$. We now divide to get $\frac{\$405}{125} = \$3.24$ per student. The average donation per student was $3.24.

You may have noticed that the values in the probability model in Example 2 can be used to find the average donation as well. $AverageDonation = (\$1)(0.64) + (\$5)(0.2) + (\$10)(0.16) = \$3.24$. Simply multiply the amount of the donation by the probability of that donation for each amount and add those results together. This leads us to our expected value formula which is given in Figure 3.1 below.

$$EV = (\text{Value } 1)(\text{Prob. } 1) + (\text{Value } 2)(\text{Prob. } 2) + (\text{Value } 3)(\text{Prob. } 3) + \ldots$$

Figure 3.1

Example 3

What is the expected value for the total of a roll for two 6-sided dice?

Solution

We will address this two ways. The first method will be done by using averaging and the second method will be done by using the expected value formula.

Begin by building the sample space for the sum of two dice. As in section 1.1, we get the dice chart shown below. Notice that there are exactly 36 equally likely spaces on the grid. So instead of playing just one time, suppose we play 36 times. If everything matches the theoretical probabilities, each of these outcomes would happen exactly one time. Add the values for each of the 36 spaces and divide by 36. For simplicity, we will add diagonally to get $\frac{2+3+3+4+4+4+...+10+10+10+11+11+12}{36} = \frac{252}{36} = 7$. The expected value is 7 which means that 7 is the average value of a roll of two 6-sided dice.

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Let's now do this problem by building a probability model and using the expected value formula, $EV = (value\ 1)\,(prob.\ 1) + (value\ 2)\,(prob.\ 2) + (value\ 3)\,(prob.\ 3) + ...$

The probability model for this situation is given below. Does this make sense and did you verify that the total of the probabilities in the table add up to 1?

Value	2	3	4	5	6	7	8	9	10	11	12
Prob.	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Using our expected value formula we have

$EV = (2)\left(\frac{1}{36}\right) + (3)\left(\frac{2}{36}\right) + (4)\left(\frac{3}{36}\right) + (5)\left(\frac{4}{36}\right) + (6)\left(\frac{5}{36}\right) + (7)\left(\frac{6}{36}\right) + (8)\left(\frac{5}{36}\right) + (9)\left(\frac{4}{36}\right) + (10)\left(\frac{3}{36}\right) + (11)\left(\frac{2}{36}\right) + (12)\left(\frac{1}{36}\right) = 7$. The expected value for the total when two dice are rolled is 7.

Example 4

A carnival game is being played that has several prizes that a player can win. Suppose a probability model is put together for this game as shown below.

a) Find the missing value.

b) Calculate the expected value and explain what it means.

Table 3.1: **Probability Model for Carnival Game**

Value	$30	$20	$10	$1
Probability	0.01	0.03	???	0.9

Solution

a) The probabilities in a probability model must add up to 1. We recognize that 0.01+0.03+???+0.9=1 must be true. The missing value must be 0.06.

b) $EV = (\$30)(0.01) + (\$20)(0.03) + (\$10)(0.06) + (\$1)(0.9) = \$0.30 + \$0.60 + \$0.60 + \$0.90 = \$2.40$. Our expected value is \$2.40. In other words, if this game were played many times, the average payout would be \$2.40. Note that the expected value of \$2.40 is not a possible prize that a player can win.

Example 5

Suppose a casino game has an expected payout of \$1 every time it is played. A player is paid nothing 45% of the time, they are paid one dollar 35% of the time, and they are paid three dollars 15% of the time. There is one more payout amount in this game.

a) Build a probability model for this situation. Be sure to calculate the percent of time the remaining payout occurred.

b) How much should this payout be so that the expected value is \$1?

Solution

a) Start by noticing we have used 45%+35%+15%=95% of all outcomes. This means that the remaining outcome must be 5%. This allows us to build a probability model that is mostly complete.

Table 3.2: **Probability Model**

Amount	\$0	\$1	\$3	???
Probability	0.45	0.35	0.15	0.05

Our calculation is now based upon the expected value formula. Use 'x' to represent the missing amount.

$(\$0)(0.45) + (\$1)(0.35) + (\$3)(0.15) + (x)(0.05) = \$1.$

$\$0 + \$0.35 + \$0.45 + (x)(0.05) = \$1$

$\$0.80 + (x)(0.05) = \$1.$

$(x)\,(0.05) = \$0.20$

x=$4. The missing value is $4.

Example 6

A carnival game has prizes and probabilities as shown in the table below. How much should the game cost if the owner of the game wants to average a $2 profit per player?

Value	$3	$5	$20
Prob.	0.65	0.30	0.05

Solution

First calculate the expected value to get $EV = \$3 \times 0.65 + \$5 \times 0.30 + \$20 \times 0.05 = \$4.45$. This means that the average player will be paid $4.45 when they play. Therefore, the owner should charge $2 more than this or $6.45.

Problem Set 3.1

Exercises

1) The table below represents the number of vehicles and the associated probability of having that number of vehicles in an individual household. What is the expected number of vehicles in a typical household?

Table 3.3: **Vehicle Ownership**

# Owned	0	1	2	3	4	5
Probability	0.02	0.26	0.37	0.19	0.12	0.04

2) A student sells products as part of a fundraiser to raise money for a choir trip to New York. She sold 75 items total which included 50 rolls of cookie dough for $6 each, 15 packages of butter braids at $10 each, and 10 bake-at-home bread packs for $12 each.

a) Find the percent of her sales for each item.

b) Build a probability model for this situation.

c) Find the expected value of a sale for this particular student.

3) The owner of Friendly's Casino decides that she will set up her payouts in their 'Fast Cash' game so that the average gambler neither wins nor loses money. For a gambler who plays this game, the chance of

getting paid nothing is 30%, the chance of getting paid $5 is 40%, the chance of getting paid $10 is 25%, and the chance of getting paid $30 is 5%. How much will the owner of Friendly's charge for this game?



4) The owner of Greedy's Casino decides he wants to make an average of $1.50 every time a gambler plays the game called 'Funny Money'. The chance of getting paid $2 is 20%, the chance of getting paid $5 is 40%, the chance of getting paid $10 is 30%, and the chance of getting paid $15 is 10%. What should the owner of Greedy's charge to play this game?

5) In a certain racing video game, players try to go around a track as many times as possible. If a racer completes a lap in time, they continue on to the next lap. If they don't complete a lap in time, their race is complete at the end of the lap they are currently finishing. The probability model below gives the probabilities of the maximum number of laps completed by people who play the video game. What is the expected number of laps completed for each racer?

Table 3.4: **Laps Completed During Racing**

# of Laps	1	2	3	4	5
Probability	0.29	0.38	0.17	0.11	0.05



6) In a certain casino game, the average payout (expected value) for a player is \$2.53. A partially completed probability model for this game is given below.

Table 3.5: **Casino Game Payouts**

Amount Paid	\$0	\$1	\$3	???	\$21
Probability	0.32	0.47	0.08	0.07	0.06

a) What is the missing amount?

b) If the casino was going to set a price for this game, do you think they would choose to charge \$2 to play or \$3 to play? Explain your choice.

c) If the casino was going to set a price for this game, do you think they would choose \$3 to play or \$6 to play? Explain your choice.

7) What is the average roll for a single 6-sided die?

8) A coin is flipped one time. If it lands on heads, you win \$20. If it lands on tails, you win \$30. Build a probability model and calculate the expected value for this game by using the expected value formula.

9) Two students are given the partially completed probability model below as part of a project. The teacher tells them that the expected value for this situation is \$6.95.

Table 3.6: **Probability Model Given to Students**

Value	\$3	\$6	\$10	\$50
Probability	0.25	0.35	???	0.07

a) Assuming that the expected value of \$6.95 is correct, what should be the value of the missing probability? Explain why this is impossible.

b) Assuming the expected value of \$6.95 is incorrect, what should be the value of the missing probability?

c) The given expected value of \$6.95 was incorrect. What is the correct expected value?

10) I want to come up with a game that has 5 prizes. There will be a 20% chance of getting paid $1, a 25% chance of getting paid $3, a 15% chance of getting paid $4, and a 30% chance of getting paid $7.

a) What is the probability of winning the 5th prize?

b) What is the value of the 5th prize if the expected payout for the game is $4.75?

11) For Halloween next year, I have decided that I will distribute an average of 1.6 pieces of candy per child who comes to my door. To help me do this, I have set up a game of chance whereby each trick-or-treater gets to play a game that determines how many pieces of candy they get to pick from my bag. I started building a probability model that shows the probabilities of being able to select 0, 1, 2, 3, or 4 pieces of candy. Unfortunately, I did not have time to finish my table. Use what I have so far to answer the questions.

Table 3.7: **Halloween Candy Distribution**

# of Pieces	0	1	2	3	4
Probability	0.03	0.45	x	y	0.02

a) Give the expected value equation using the variables x and y and the expected value of 1.6. Simplify your equation by combining like terms.

b) Give an equation using the variables x and y that uses the fact that the probabilities in a probability model must add up to one. Clean up your equation by combining like terms.

c) Using your answers from parts a) and b), write a system of equations and solve for the variables x and y.

Review Exercises

12) A sample of 325 students were asked which electronic device they use most frequently, their cell phone, a computer (including wireless devices), or a television (including video games). The gender of the student was also recorded and the results are shown in the two-way table below.



	Cell Phone	Computer	Television	Total
Male	60	30	55	145
Female	115	45	20	180
Total	175	75	75	325

a) What is the probability that a randomly selected student was a male?

b) What is the probability that a randomly selected student was most likely to say they used a cell phone most frequently?

c) What is the probability that a student was male given that they indicated they used the television most frequently?

d) What is the probability that a student indicated that they used a computer most frequently given that they were a female?

13) One floor of an office building is being remodeled and redecorated and an employee is responsible for picking out three different styles of chair and two different styles of table for their office furniture. Suppose the furniture store has 10 different chair styles and 4 different table styles that would be appropriate for office furniture. In how many different ways can the employee select the three chair styles and two table styles?

14) Consider a standard deck of 52 cards. Suppose 1 card is drawn randomly from the deck. Find each probability.

a) P(Red Card)

b) P(Spade)

c) P(Face)

d) P(Heart|Red)

3.2 Applied Expected Value Calculations



(Note- Three Separate Videos for the Section)

Learning Objectives

- Understand the concept of a fair game
- Be able to analyze a game of chance by building a probability model and calculating expected values from scratch

Casinos have a very delicate balancing act they must manage. First of all, they want to make money. However, people just don't like to lose money. In order to make money, the casino games have to be in favor of the house and not the player. Why don't casinos tilt the games even more to the house's favor? If they did, their expected value would certainly go up. On the other hand, attendance at the casino would go down.



No matter what the odds, a casino can't make money unless they can keep people coming through the doors. Setting the games up so that there are still winners, some occasionally big, is good for attendance. You might even know someone who has made a large amount of money at a casino.

In this section, we will bring together our ideas about calculating probabilities from Ch. 2 along with the concept of expected value in order to be able to analyze a game of chance. We begin where we left off in section 3.1. A **fair game** is a game in which neither the player nor the house has an advantage. In other words, when all is said and done, the average player will not have made or lost any money whatsoever.

Example 1

A bag has 10 red marbles and 8 blue marbles in it. A player reaches into the bag pulling out 2 marbles, one after the other without replacement. If the color of the two marbles match, the player wins $10. If they don't match, the player wins nothing. The game costs $5 to play.

a) Use a tree diagram to help find the probability that the two marbles match. Use your result to build a probability model.

b) Is this game a fair game? If so, explain why. If not, give the value that the game should cost in order to be fair.

Solution

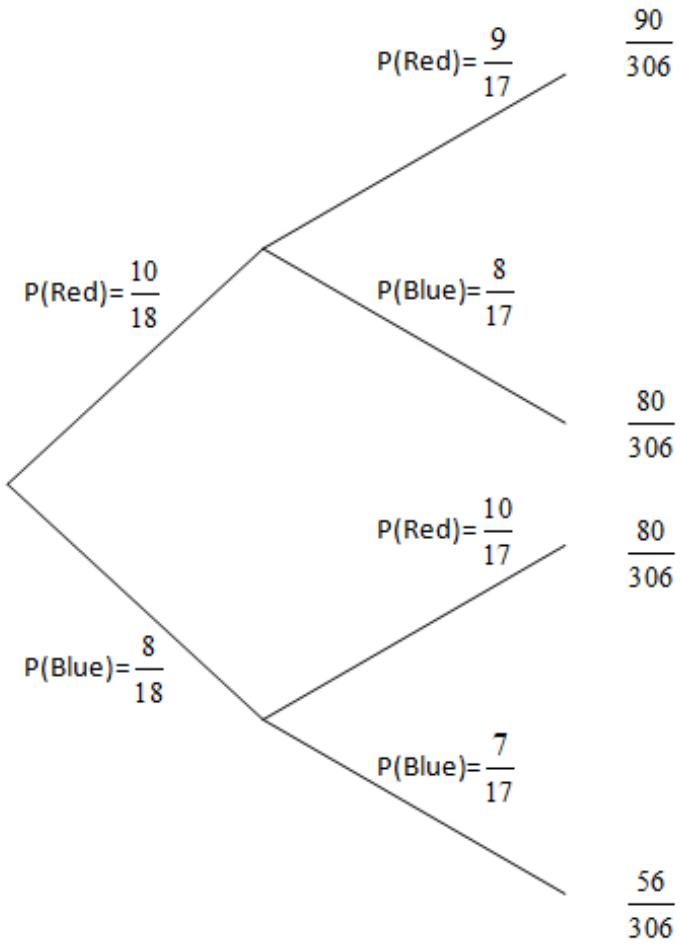


Table 3.8: **Marble Distribution**

Result	Red,Red	Red,Blue	Blue,Red	Blue,Blue
Value	$10	$0	$0	$10
Probability	$\frac{90}{306}$	$\frac{80}{306}$	$\frac{80}{306}$	$\frac{56}{306}$

$$EV = (\$10)\left(\frac{90}{306}\right) + (\$0)\left(\frac{80}{306}\right) + (\$0)\left(\frac{80}{306}\right) + (\$10)\left(\frac{56}{306}\right) = \frac{\$1{,}460}{306} \approx \$4.77$$

The expected value is \$4.77. Notice, however, that \$4.77 is the expected amount that the house *pays out* each game. The expected value for the house is \$0.23 because every player must pay \$5 to play. At \$5, the game is not fair because it favors the house by an average of 23 cents every time the game is played. To be a fair game, it should cost \$4.77.

The game of *GREED* is a game of chance in which players try to decide when they have accumulated enough points on a turn to stop. Two 6-sided dice are rolled. The player gets to keep the total that shows on the two dice. After every roll, the player can either decide to roll again and try to add to their current total for that turn or stop and put their points in the bank. The only catch in this game is that if a total of 5 is rolled, all points accumulated on that turn are lost. For example, suppose the first roll has a total of 9 and the player decides to go again. The next roll has a total of 7. The player now has 16 points accumulated on this turn and must decide to either put those 16 points in the bank or risk them. If they decide to risk the 16 points and a total of 5 comes up next, the score for that turn will be 0.

Example 2

Suppose a person is playing *GREED* and has accumulated 26 points so far. Is it to their advantage to roll one more time?

Solution

We will build a probability model and calculate the expected value based upon what might happen with one more roll. (See Example 3 from Section 3.1.) For example, there is a $\frac{1}{36}$ chance that the total will be 2. This would mean the player would have a total of 28 points with one more roll. The highest a player could have after this turn would be 38 points if they happen to roll a total of 12. The risk is that the player will roll a total of 5 and lose their 26 points.

Value	28	29	30	0	32	33	34	35	36	37	38
Prob.	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The one item to be careful about here is that it is impossible to get a total of 31 in the chart. Remember, if you roll a total of 5, you lose all of your points. When we perform our expected value computation for the probability model above, we get approximately 29.6. In other words, if we roll exactly one more time, our average result will be almost 30 points. This is definitely better than stopping with 26 points. It is to the advantage of the player to roll again.

Example 3

An investor is going to make a long-term investment in a company. If all goes well, an investment of \$100 will be worth \$900 in twenty years. The risk is that the company may go bankrupt within twenty years in which case the investment is worthless. Suppose there is a 25% chance that the company will go bankrupt within 20 years. What is the expected value of this investment?

Solution

Start by building a probability model as shown below that shows that there is a 25% chance of making nothing and a 75% chance of making $900.

Value	$0	$900
Prob.	0.25	0.75

$EV = (\$0)\,(0.25) + (\$900)\,(0.75) = \$675$. Taking into account that this investment cost $100, the investor should get an average profit of $575.

Problem Set 3.2

Exercises

1) In the carnival game W*iffle Roll,* a player will roll a wiffle ball across some colored cups. Suppose that if the ball stops in a blue cup, the player wins $20. If it stops in a red cup, they win $10, and if it stops in a white cup, the player wins nothing. There are 25 white cups, 4 red cups, and 1 blue cup. Assume the chances of stopping in any cup is the same. How much should this game cost if it is to be a fair game?

2) In a simple game, you roll a single 6-sided die one time. The amount you are paid is the same as the amount rolled. For example, if you roll a one, you get paid $1. If you roll a two you get paid $2 and so on. The only exception to this is if you roll a 6 in which case you get paid $12. What should this game cost in order to be a fair game?

3) In the Minnesota Daily 3 lottery, players are given a lottery ticket based upon 3 digits that they pick. If their 3 digits match the winning digits in the correct order, then the player wins $500. If the digits don't match, then the player loses. The game costs $1 to play. What is the expected value for a player of this lottery game.



4) Suppose you are playing the game of *GREED* as described in Example 2. You have accumulated a total of 55 points on one turn so far. Is it to your advantage to roll one more time?

5) Suppose you are playing the game of *GREED* again. This time you have accumulated a total of 60 points in one turn so far. Is it to your advantage to roll one more time?

6) Using your results from numbers 4) and 5) and a little more investigation, for what number of points in a turn in the game of *GREED* does it make no difference if you roll one more time or stop? In other words, at what point total does the expected value with one more roll give the same total as if you had stopped?

7) A bucket contains 12 blue, 10 red, and 8 yellow marbles. For $5, a player is allowed to randomly pick two marbles out of the bucket without replacement. If the colors of the two marbles match each other, the player wins $12. Otherwise the player wins nothing. What is the expected gain or loss for the player?

8) An insurance company insures an antique stamp collection worth $20,000 for an annual premium of $300. The insurance company collects $300 every year but only pays out the $20,000 if the collection is lost, damaged or stolen. Suppose the insurance company assesses the chance of the stamp collection being lost, stolen, or destroyed at 0.002. What is the expected annual profit for the insurance company?

9) A prospector purchases a parcel of land for $50,000 hoping that it contains significant amounts of natural gas. Based upon other parcels of land in the same area, there is a 20% chance that the land will be highly productive, a 70% chance that it will be somewhat productive, and a 10% chance that it will be completely unproductive. If it is determined that the land will be highly productive, the prospector will be able to sell the land for $130,000. If it is determined that the land is moderately productive, the prospector will be able to sell the land for $90,000. However, if the land is determined to be completely unproductive, the prospector will not be able to sell the land. Based upon the idea of expected value, did the prospector make a good investment?

10) A woman who is 35 years old purchases a term life insurance policy for an annual premium of $360. Based upon US government statistics, the probability that the woman will survive the year is .999057. Find the expected profit for the insurance company for this particular policy if it pays $250,000 upon the woman's death.

11) A bucket contains 1 gold, 3 silver, and 16 red marbles. A player randomly pulls one marble out of this bag. If they pull a gold marble, they get to pick one bill at random out of a money bag containing a $100 bill, five $20 bills, and fourteen $5 bills. If they pull a silver marble out of the bag, they get to pick one bill at random out of a bag containing a $100 bill, two $20 bills, and seventeen $2 bills. If your marble is red, you automatically lose. The game costs $5 to play.

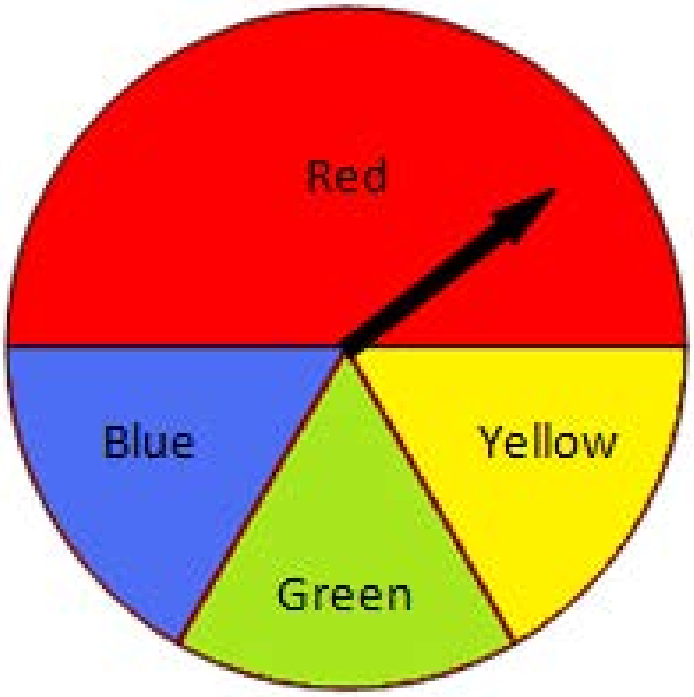
a) Build a tree diagram for this situation.

b) Build a probability model for this situation.

c) Calculate the expected gain or loss for the player.



12) A spinner has four colors on it, red, blue, green, and yellow. Half of the spinner is red and the remaining half of the spinner is split evenly among the three remaining colors. A player pays some money to spin one time. If the spinner stops on red, the player receives $2. If it stops on blue, the player receives $4. If it stops on either green or yellow, the player wins $5. What should this game cost in order to be a fair game?



13) A bag contains 1 gold, 3 silver, and 6 red marbles. A second bag contains a $20 bill, three $10 bills, and six $1 bills. A player pulls out one marble from the first bag. If it is gold, they get to pick two bills from the money bag (without replacement). If it is a silver marble, they get to pick one bill from the money bag, and if the marble is red, they lose. The game costs $3 to play. Should you play? Explain why or why not.

14) Suppose the Minnesota Daily 3 lottery adds a new prize. You still get $500 if you match all three digits in order, but you can also win $80 if you have the three correct digits but not in the right order. The game still costs $1. What is the expected value for a player of this lottery game?

Review Exercises

15) Consider the partially complete probability model given below.

Value	$2	$4	$7	$11
Prob.	X	.25	.15	.05

a) What is the value of 'X'?

b) What is the expected value for this situation.

16) The student council is starting to prepare for prom and decides to name a committee of 6 members. Suppose that they decide the committee will have 2 juniors and 4 seniors on it. In how many ways can the committee be selected if there are 8 juniors and 8 seniors from which to select?

17) Three cards are dealt off the top of a well-shuffled standard deck of cards. What is the probability that all three cards will be the same color?":

18) A student does not have enough time to finish a multiple choice test so they must guess on the last two questions. List the sample space of the possible guesses for the last two questions if each question has only choices a, b, and c.

3.3 Simulation and Experimental Probability



(Note- Two Separate Videos for the Section)

Learning Objectives

- Understand how to generate random numbers using a random digit table or a calculator
- Be able to properly assign digits to simulate a random situation
- Be able to interpret results from a simulation and understand the connection to the Law of Large Numbers

For many of the problems we have addressed, putting together a theoretical model is very reasonable for us to do. A **theoretical model** gives a picture of what *should* happen in the long run for any situation involving probability. It will give us a very clear idea of what to expect out of a particular situation. If you have been dealt an ace, you can quickly figure out the probability that the next card will be a face card. That probability is a theoretical probability.

However, the truth is that in many situations it is beyond the scope of the mathematics of this course to calculate theoretical probabilities. In these situations, we can estimate probabilities by performing a **simulation** through the use of an experimental model. Some of our simulations can be done quite easily using actual probability tools like dice or spinners. Some situations, though, will require us to use a **random number generator** or a **table of random digits**. Be sure you can find the random number generator on your calculator.

You can see a table of random digits at the end of this book in Appendix A, Part 1. A table of random digits contains a random mix of digits from 0 through 9. These digits can be used to simulate virtually any situation involving chance behavior from rolling dice to drawing cards to modeling other known situations. It is important that you are detailed in your explanation of how you will assign digits so that others may model your simulation procedure exactly. The 6 steps for using a table of random digits for a simulation are shown below.

1) Assign the same number of digits to each of the different possible outcomes.

2) Choose a line number to use from the random digit table. Sometimes the line number is given to you.

3) State how many digits you will select at a time. If your largest value is less than 10, you will be able to select one digit at a time. If it is less than 100 you will have to use two digits at a time. If it is less than 1000, you will have to use 3 digits at a time and so on.

4) State what values you will ignore. These typically are values that are larger than your biggest value and number combinations like 000.

5) Know when to stop. Pay attention to the number of trials you must complete.

6) Select your digits and summarize your results.

Example 1

Use the line of random digits below to randomly select three days from the month of October.

3 5 4 7 6 5 5 9 7 2 3 9 4 2 1 6 5 8 5 0 0 4 2 6 6 3 5 4 3 5 4 3 7 4 2 1 1 9 3 7

Solution

October has 31 days in it so we must identify 31 different outcomes. Our largest value is 31 so I will have to select two digits at a time. I will ignore 32 through 99 and 00. Assign two digits per day, so for example, 01 = the first of October.

~~3 5~~ ~~4 7~~ ~~6 5~~ ~~5 9~~ ~~7 2~~ ~~3 9~~ ~~4 2~~ 1 6 ~~5 8~~ ~~5 0~~ 0 4 2 6 6 3 5 4 3 5 4 3 7 4 2 1 1 9 3 7

Notice that we crossed out 35, 47, etc... because these were all beyond our largest value of 31. Our three days are the 16th, 4th, and 26th of October.

Example 2

Suppose you wish to roll two dice a total of 5 times and keep track of the totals. You don't have any dice, but you do have access to the line of random digits below. Explain how you could simulate the rolls of two dice using the random digits and then perform the simulation.

19223 95034 05756 28713 96409 12531 42544 82813

Solution

We will have to roll two dice. Each die will have a value from 1 through 6 on each roll. We will ignore digits 7, 8, 9, and 0 as a die can never come up with those values. The first three digits in the line of random digits are 1,9,2. The first die will be a 1. We ignore the 9. The second die will be a 2. This gives a total of 3. Our second roll picks up right where we left off and we get a 2 and a 3 for a total of 5. Using the same procedure, our next three totals will be 8, 9, and 11. Our five results were 3, 5, 8, 9, and 11.

Remember that it is unwise to make assumptions after only a very small set of rolls. For example, it would be incorrect to say that a total of 7 is unlikely to happen since it did not come up on our simulation. We only simulated the rolls 5 times which is not nearly enough to make a conclusion. The **Law of Large Numbers** states that as we increase the number of trials we should get closer and closer to the theoretical probability. Theoretically, there is a $\frac{1}{6}$ chance that the total is 7. If we did our simulation for thousands of rolls, we would expect that our simulation would show that a total of 7 comes up about $\frac{1}{6}$ of the time.

Example 3

At the start of this season, Major League Baseball fans were asked which American League Central team would be most likely to win the division this year. The table below gives the results of the poll.

Table 3.9: **Most Likely to Win AL Central**

Team	Chicago	Cleveland	Detroit	Kansas City	Minnesota
Probability	0.14	0.23	0.33	0.02	0.28

Using the line of random digits supplied, simulate the results when asking 10 fans who they think will win the AL Central.

73676 47150 99400 01927 27754 42648 82425 36290



Solution

Notice that the probability adds up to 100%. Selecting two digits at a time works well in situations like this. Since 14% or 14 out of 100 fans support Chicago, assign 01-14 as Chicago fans. Assign 15-37 as fans who think Cleveland will win. A good trick to remember is to add the 23% for Cleveland to the ending percent for Chicago which was 24%. This will give you the end of the interval for Cleveland. Likewise, Detroit will be 38-70, Kansas City will be 71-72, and Minnesota will be 73-99 and 00. Notice that Minnesota can't use 100 as that is three digits and we are only selecting two at a time. The digit combination '00' can be used to represent 100.

We now select our 10 fans. Our 2-digit pairs are 73, 67, 64, 71, 50, 99, 40, 00, 19, and 27. The table below summarizes our results.

Team	Chicago	Cleveland	Detroit	Kansas City	Minnesota
Values	01-14	15-37	38-70	71-72	73-99, 00
# of Fans	0	2	4	1	3

In our simulation, we found that 4 out of our 10 randomly selected fans felt Detroit was going to win. While it is not exactly the 33% we were given in the original problem, it is fairly close. Once again, if we had done hundreds of trials instead of just 10, our percentages would tend to get very close to the theoretical probability according to the Law of Large Numbers.

Example 4

Every person is born on a different day of the month. Some people are born on the 1st and some people are born as late as the 31st. How many people must you go through until you find two that were born on the same day of the month? Simulate this one time using the random digits below. (Ignore the fact that people are not equally likely to be born on all days. It is more likely you were born on the 17th than the 31st since all months have a 17th but not all months have a 31st.)

45467 71709 77558 00095 32863 29485 82226 90056

52711 38889 93074 60227 40011 85848 48767 52273

Solution

We will select 2 digits at a time as our largest value, 31, requires two digits. We will use 01, 02, ... 30, 31 and ignore 32-99 and 00.

The numbers we get are 45, 46, 77, 17, 09, 77, 55, 80, 00, 95, 32, 86, 32, 94, 85, 82, 22, 69, 00, 56, 52, 71, 13, 88, 89, 93, 07, 46, 02, 27, 40, 01, 18, 58, 48, 48, 76, 75, 22, and 73. The only 'keepers' are 17, 09, 22, 13, 07, 02, 27, 01, 18, and 22. We did not get a match until we got our second 22. It took us 10 people to find a pair that were born on the same day of the month.

Notice that when you get to the end of one line in a random digit table, you simple continue by moving to the next line below.

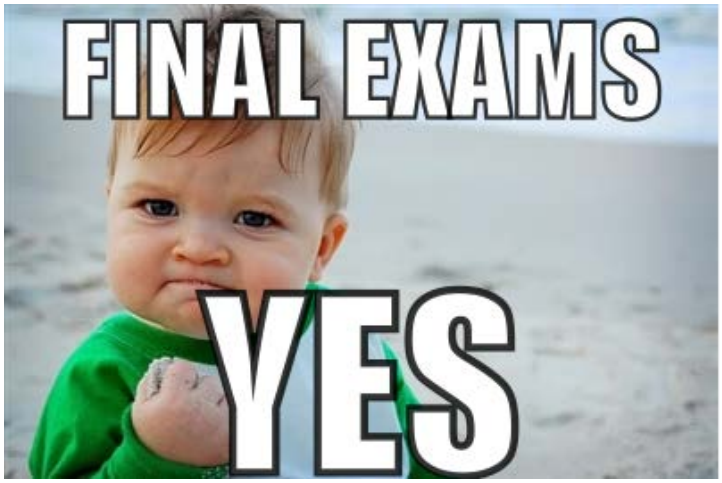
Problem Set 3.3

Exercises

1) Suppose that 80% of a school's student population is in favor of eliminating final exams.

a) Explain how could you assign digits from a random digit table to simulate this situation?

b) Suppose you ask 10 students if they would like to eliminate final exams. Simulate a random selection of 20 students and record how many of the 20 are in support of eliminating final exams. Use line 147 from the random digit table in Appendix A, Part 1.



2) Suppose that students at a particular college are asked about their class rank when they were in high school. The table below shows what they said.

Table 3.10:

Class Rank	Top 10%	Top 10% to 25%	Top 25% to 50%	Bottom 50%
Prob.	0.2	0.4	0.3	???

a) What must the probability be for the bottom 50%?

b) Explain how you could assign digits to carry out a simulation for this situation.

c) Using your set up, perform a simulation. Use 20 students in your simulation and record your results. Use line 103 from the random digit table.

3) Suppose the grades for students in your Stats & Prob. course were distributed as shown in the table below.

Table 3.11:

Grade	A	B	C	D or F
Prob.	0.20	0.29	0.35	0.16

a) Explain how you could assign digits to simulate the grades of randomly chosen students.

b) Simulate the grades for 30 students. Use line 106 from the random digit table. Build a tally chart to track your results.

c) How closely did your simulation match the actual distribution?

4) How many five card poker hands must you be dealt in order to get a hand with two cards that have matching values? (For example, the 7 of hearts and 7 of diamonds have matching values.)

a) Explain how you will assign digits for this situation.

b) Perform the simulation one time and state how many five-card hands it took for you to get your first hand with two cards that match. Use line 138 from the random digit table.



5) There are some basic concepts that should be clearly understood about a random digit table. Answer the questions below.

a) Is it possible to have four 6's next to each other in a random digit table?

b) What percent of the digits in a random digit table are 9's?

c) What should you do if you come to the end of a line of random digits and you still need more digits?

6) Suppose we have a class of 30 students and you are wondering what the chances are that there is at least one pair of students who have the same birthday. Assume that there are 365 days in a year.

a) Explain how could you assign digits from a random digit table to simulate this situation?

b) Perform this simulation one time and record whether or not there was a match in the class of 30 students. Use line 121 from the random digit table.

Review Exercises

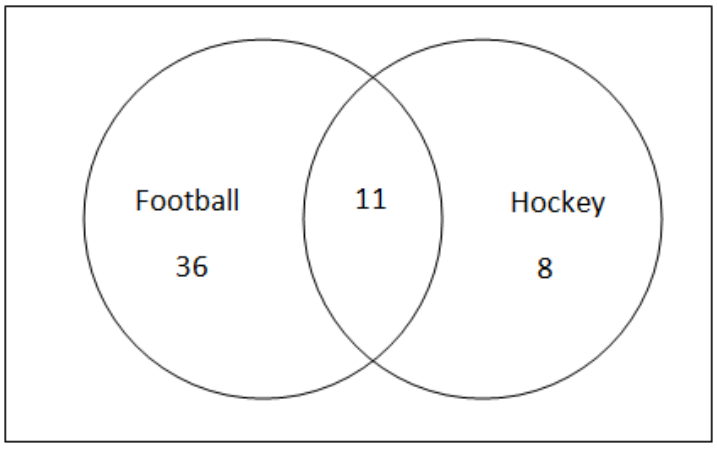
7) Suppose you are dealt two cards from a well-shuffled standard deck of 52 cards. What is the probability that your two cards are a king and an ace (in either order)?

8) Consider a set of 15 pool balls. Pool balls numbered 1 - 8 are solid and pool balls numbered 9 - 15 are striped. You pull two pool balls randomly out of a bag without replacement.

a) What is the probability that your second pool ball will be solid if your first pool ball had an even number?

b) What is P(Even|Striped)?

9) A survey of junior boys found that 97 are planning to participate in either a fall sport or in a winter sport or both. Use the Venn Diagram below to answer the questions.



a) How many junior boys are planning on playing hockey or football (or both)?

b) How many junior boys who are planning on participating in a fall or winter sport are not represented in the Venn Diagram?

c) What is P(Hockey|Football)?

3.4 Chapter 3 Review

The expected value gives us the average result over the long term. We use expected value tables and the simple formula $EV = (Value1)(Prob1) + (Value2)(Prob2) + ...$ to calculate the expected value. We can put everything together for a full probability analysis of a situation by using our probability calculations and other tools like a tree diagram. Casinos are cognizant of what the expected value is on any of their games and are confident, despite having to occasionally give away some substantial prizes, that their games will make them money in the long run. We cannot ever predict with certainty what is going to happen in a given situation, but we can always run a simulation to approximate what can happen. We will often use a random number generator or a table of random digits to help us run a simulation.

Chapter 3 Review Exercises

1) Ten red marbles and 15 blue marbles are in a bag. A game is played by first paying \$5 and then picking two marbles out of the bag without replacement. If both marbles are red you are paid \$10. If both marbles are blue, you are paid \$5. If the marbles don't match, you are paid nothing. Analyze this game and determine whether or not it is to your advantage to play.

2) When two dice are rolled, you can get a total of anything between 2 and 12.



a) Use the table of random digits in Appendix A, Part 1 to simulate rolling two dice 36 times. Begin on line 119. Make a chart displaying the different results that you get and how many times you get each result.

b) How close was your simulation to the theoretical probability of what should happen in 36 rolls?

3) A bag contains a \$100 bill and two \$20 bills. A person plays a game in which a coin is flipped one time. If it is heads, then the player gets to pick two bills out of the bag. If it is tails, the player only gets to pick one bill out of the bag. How much should this game cost to play if it is to be a fair game?

4) Suppose there are 38 kids in your Statistics and Probability class. Devise a system using a random digit table so that the teacher can randomly select 4 students to each do a problem on the board. Use line 137 from the random digit table to carry out your simulation and state the numbers of the four students who are selected.

5) A spinner with three equally sized spaces on it are labeled 1, 2, and 3. A bag contains a $1 bill, a $5 bill, and a $10 bill. A player gets paid the amount they pull out of the bag times the number that they spin. What should this game cost in order to be a fair game?

6) The table below shows the probabilities for how kids get to school in the morning.

Table 3.12: **How students get to school**

Method	Bus	Walk	Car	Other
Probability	0.31	0.14	0.39	???

a) What must the *Other* category have as a probability?

b) Describe how you would assign digits from a random digit table to set up a simulation for selecting a student to find out how they got to school.

c) Carry out your simulation for a total of 10 students and record your results. Use line 104 from the random digit table.

7) In a sharpshooting competition, competitors shoot at a total of 20 targets. The table below shows the probabilities associated with hitting a certain number of targets. Some shooters are perfect but even the poorest shooters still hit at least 15 targets.

a) What is the most likely number of hits that a shooter will have?

b) What is the expected number of hits that a shooter will have?

Table 3.13: **Shooting Accuracy out of 20 shots**

# of Hits	15	16	17	18	19	20
Probability	0.04	0.12	0.35	0.28	0.18	0.03

8) In a game of chance, players pick one card from a well-shuffled deck of 52 cards. If the card is red, they get paid $2. If the card is a spade they get paid $3. If the card is a face card, they get paid $5 and if the card is an ace they get paid $10. A player gets paid for all the categories they meet. For example, the King of Spades would be worth $8 because it is a spade and a face card. How much should this game cost in order to be a fair game?

Image References

Slot Machine http://www.gamedev.net

Welcome to Las Vegas http://pilipon.wordpress.com

Race Cars http://www.thunderboltgames.com

Electronic Devices http://www.topnews.in

Poker Chips http://www.ppppoker.com

Minnesota State Lottery http://www.mnlottery.com

$100 Bills http://www.sciencebuzz.org

MN Twins Logo www.twins.mlb.com

Final Exams Yes http://www.york.org

Pair of Jacks http://xdeal.com

Dice http://goblin-stock.deviantart.com

Chapter 4

Data Collection

Introduction

What is data? Why collect data? How is data collected? Who cares anyway?

- *How many walleye are in Lake Mille Lacs?*
- *Does aspirin prevent heart attacks?*
- *What is the approval rating for the President?*
- *How have the schools in Minnesota been doing to prepare students for success in college?*
- *What percent of people would return money when given too much change?*

All of these questions (and infinitely many more) can be answered through statistics. Statisticians begin by posing a question. Then they plan a method for collecting information, called data, about that question. Next they collect the data and analyze it. The statisticians will 'look' at the data in the form of graphs or tables. They will 'analyze' the data with numerical statistics. Finally they will 'explain' what they have learned, what conclusions can be made, and what is still unknown, in a written or verbal report.

4.1 DATA

Learning Objectives

- Know the terminology of data collection, variables, and measurement
- Understand how measurements are used in statistics
- Distinguish between the various methods for data collection

Data and Variables

When a topic needs to be studied or a question needs to be answered, researchers often collect data in an effort to find the answer. **Data** is a collection of facts, measurements, or observations about a set of individuals (data is plural, the word datum refers to a single observation). There are a variety of ways to collect data in order to study topics of interest. Researchers can analyze and compare test scores for various Minnesota High Schools. Scientists can conduct an experiment to determine the effectiveness of

a new medication. Union leaders can conduct a census of every union member before deciding to strike. Or market researchers can survey a randomly selected sample of teenage girls to determine what qualities they look for when purchasing a new cell phone.

When a topic is being studied, there are often several **variables,** or characteristics about the individuals, that the researchers are interested in. Each person, animal, or object being studies is one **individual** (or subject). The variables can be either categorical or numerical. A **categorical variable** (or qualitative variable) can be put into categories, like favorite colors, type of car, etc. Whereas a **numerical variable** (or quantitative variable) can be assigned a numerical value, such as heights, distances, temperatures, etc.



Associated Press / Kristin Murphy

Example 1

Suppose 1845 teenage girls are to be surveyed by a cell phone company that wants to design a new cell phone that they can successfully market to females under 20 years old. The questionnaire will likely include questions related to: age, birth date, race, area code where they live, number of texts sent per month, amount of money willing to spend per month, services they want offered, features they want included, length of time they have had a cell phone, favorite colors, etc. All of these are **variables**, because they will vary from individual to individual. However, only some of these variables are numerical. Identify the individuals and the numerical variables.

Solution

Individuals: each girl who completes a questionnaire

Numerical Variables are: age, number of texts per month, amount of money willing to spend per month, length of time they have had a cell phone.

When determining which variables are numerical, it might help to decide whether or not it would be appropriate to calculate an average, or the range for the reported data. Age is numerical, because we can certainly report an average age of those surveyed. Even though birth date and area code may be reported as numbers, it would make no sense to report an 'average birth date' or 'mean area code'. Numbers such as these, social security numbers, or student ID numbers divide the data into a bunch of categories of one item each. They are simply used for identification and are considered categorical variables.

Measurement in Statistics

When a topic is to be studied the researchers decide what it is they want to know about each individual. These **variables of interest** can be measured using different instruments and need to be reported in specific units. The **instrument** is the tool used to make the measurement. This instrument could be something obvious like a scale, tape measure, thermometer, or speedometer. But, it could also be a something like a questionnaire, a rubric, or an exam. The **units** explain what the numbers represent, and might be feet, points, pounds, degrees Celsius, miles per hour, etc. Keep in mind that data is useless unless it is in context. For example, the number 12 could mean anything. Is it $12, or 12 inches, or 12 (in $1,000), or 12 apple pies? Without knowing the units, all you have is a meaningless list of numbers.

Validity, Reliability and Bias

The way in which any given variable is to be measured needs to be valid and reliable. **Validity** refers to the appropriateness of the instrument and units used. **Reliability** means that the instrument can be depended upon to consistently give the same measurement (or nearly the same). If an instrument gives different results when measuring the same thing, it is not reliable, and it has a lot of variability (because the results vary a lot). Another potential problem with measurements is bias. When a measurement is repeatedly too high or too low, it is said to be biased. In other words, a **biased** measurement is 'consistently wrong in the same direction'.

Researchers would like to limit bias in measurements as much as possible. Ideally, we hope for measurements that are valid, low in bias, and highly reliable. No measurement is perfect or necessarily accurate. Averaging repeated measurements can be a way to limit variability. Be aware though, averaging will only reduce variability (or increase reliability). Averages will not make an in invalid measurement suddenly valid. And, the average of biased measurements will still be biased.

For example, if the variable being studied is the weight of all of the members of the school wrestling team, then using a *scale* as the instrument and *pounds* as the units will be valid. And, as long as the scale being used is in working order, then the measurements reported should be reliable.

However, what if someone had set the scale being used to weigh the wrestlers to start at 10 pounds rather than zero? Each person who stepped on the scale would think that they were 10 pounds heavier than they actually were, resulting in biased measurements. If that were the case, using the scale as the instrument and pounds as the units would be valid (makes sense as a way to measure weight); reliable (if the same person steps on the scale again and again, they will have nearly the same result); and biased (each measurement is 10 pounds too heavy). So, even though something is wrong with this measurement, it doesn't mean that everything is wrong with it. We want valid, reliable, and unbiased measurements.

Example 2

Suppose that a teacher intends to base grades in a math class on the students' heights. She plans to use a *tape measure* as her instrument and *inches* as her units. Her grading system will be as follows: the shortest student will receive the lowest grades and the tallest will receive the highest grades. Comment on the validity, reliability, and potential bias of this.



Solution

Validity? This clearly is not a valid way to measure a student's success and assign grades, because height has absolutely nothing to do with someone's understanding of math, or grade in a math course.

Reliability? The tape measure should be reliable. If used properly, each time a particular student's height is measured we will expect to get the same answer.

Bias? This should not be biased. Some tall people will deserve higher grades, while some will deserve lower grades. The same will be true for students of all heights.

Therefore, this teacher's method for assigning grades would be unbiased and it would be reliable (both good things), but it would also not be valid (a bad thing). She should come up with a better way to measure students' grades. Perhaps a combination of test scores and homework completion.

So, keep in mind that just because a statistical measurement is bad, does not mean that everything will be wrong with it. It is important to think through each question separately: *Is it valid?; Is it reliable?; Is it unbiased?*

Rates versus Counts

Something to watch out for is whether numbers should be changed to rates or percentages in order to make appropriate comparisons. For example, it would not make any sense to compare 'the number of people living in poverty' for each of the fifty states in the United States because of the variety in population sizes. Think of the number of people who live in the state of Rhode Island versus the number who live in California. It would be much more appropriate to compare 'the percentage of people living in poverty' for each state instead.

Example 3

Luigi got a pair of jeans that are normally $64.95, for $52.50. Javier paid $48.75 for a pair of jeans that normally cost $58.25. Which jeans had a higher rate of discount?

Solution

Luigi's jeans were marked down $12.45 (64.95 - 52.50). Divide the amount of discount by the original cost (12.45/64.95) and get 0.1917. So, Luigi's jeans were marked down 19.17%.

Javier's jeans were marked down $9.50 (58.25 - 48.75). Divide the amount of discount by the original cost (9.50/58.25) and get 0.1631. So, Javier's jeans were marked down 16.31%.

Therefore, Luigi's jeans had a higher rate of discount.

Methods for Collecting Data

Once a question of interest is posed, there are different ways of collecting data. This is a quick overview of the methods for collecting data that will be studied in this chapter: sample surveys, census, observational studies, and experiments. Each will be covered in more detail in the following sections. As of now, we just want to be able to recognize which method was used or described.

Sample surveys are often used as a way to collect data from just some of the people or objects being studied. Some examples of sample surveys are: mailed out questionnaires, online surveys, phone interviews, or quality control checks. Another way to collect data is through a **census**. This means that every single person or item in the population is checked, tested, or asked. When trying to determine whether something was a sample or a census, ask yourself if the researchers asked everyone (or tested everything). If yes, then it was a census.

Sometimes it will be most appropriate to conduct an **experiment** - when the researchers actually 'do something' to the subjects. Observational studies are another common way to collect data. In **observational studies**, the researchers do not 'do anything' to the subjects, they simply collect data that has already happened or happens naturally. All of these methods of data collection can yield interesting results and often answer questions. However, the only method that can actually prove that one variable *causes* another is an experiment. When trying to determine whether a research method was an experiment, ask yourself if the researchers did anything to the people or objects that were being studied? If yes, then it was an experiment.

Example 4

For each of the following scenarios, determine whether the situation described is an experiment, observational study, census, or a sample survey.

a) Researchers suspected that aspirin could help reduce the risk of having a heart attack. Seven hundred people, aged 40 or older, were willing to participate in a study. Half of these participants were randomly selected to take an aspirin each day. The remaining participants were given a pill that looked like the aspirin, but contained no actual medicine. The study went on for five years and the participant's health was monitored.

b) In an effort to study how the high schools in Minnesota have been preparing students for college, an extensive questionnaire was developed. Ten percent of the high school juniors at every high school in the state were selected randomly to complete this questionnaire.

c) Researchers suspected that tanning beds caused skin cancer. Each time a person was diagnosed with skin cancer, they were asked a serious of questions including whether or not they had used a tanning bed. If they had, further questions were asked regarding how often, what type, and at what age, etc.

d) In an effort to determine how many fish were in Lake George, the lake was drained and the fish were counted.

Solution

a) This is an experiment because the researchers changed something. They had the people take aspirin (or fake aspirin).

b) This is a sample survey because only a part of all of the high school students were questioned.

c) This is an observational study because no change was made. The researchers simply asked about past behavior.

d) This is a census because every fish was counted. However, this is ridiculous!! So, let's hope they can find a better way to determine how many fish are in a lake next time!

Problem Set 4.1

Section 4.1 Exercises

1) Lucas is writing an article about the baseball teams for the school paper. He collects data about each player's position, batting average, number of at-bats, hits, stolen bases and whether each player is on the junior varsity or varsity team. Who are the individuals? Which variables are categorical? Which are numerical?

2) Malia has been put in charge of analyzing the employees at her company. She collects information regarding annual salary, years with the company, highest degree earned, job title, yearly contribution toward 401K, number of children, home address and phone number. Who are the individuals? Which variables are categorical? Which are numerical?

3) Determine whether each of the following variables is categorical or numerical.

a) The heights of all of the volleyball players.

b) The position played by all of the football players.

c) The brand of mascara preferred by those surveyed.

d) The numbers of texts sent per month.

e) Each person's social security number.

f) Each person's cell phone provider.

4) The fourth graders at Sand Creek Elementary are doing a unit on weather. There is a thermometer on the building just outside the classroom window. The students will record and analyze the temperature at 8:00 a.m. and 2:00 p.m. every school day for 5 weeks, and then create a graph and write a report based on their findings.

a) Identify the variable of interest, the instrument used, and the units.

b) Comment on the validity, reliability and potential bias for this study.

5) The first graders at Sand Creek Elementary are doing a unit on measurement. Each student has traced her or his own foot and cut it out. Each student will use his or her 'foot' to measure various objects around the room and school. Some of the measurements they will make are height of self and at least two other friends, width of the classroom door, length of a lunch table, etc.

a) The variables of interest are the lengths, widths and heights of various objects. Identify the instrument used, and the units.

b) Comment on the validity, reliability, and potential bias for this study.

6) Determine whether each of the following measurements would have a problem with any of the following: VALIDITY (*problem would be a lack of)*, RELIABILITY (*problem would be a lack of*), BIAS, (*problem would be the presence of*). A measurement may have any combination of the factors. For each one with a problem, suggest a better way to make the measurement. (hint: answer similar to example #2)

a) A speedometer is totally unpredictable.

b) Cholesterol levels are determined by patients filling out a survey regarding their diet.

c) Time is measured by using the clock on a cell phone.

d) Grades in a Physics class are determined by students assessing themselves on a scale of 1 to 10.

e) Grades in a statistics class are determined by students' scores on one cumulative test.

f) Sobriety is determined by a breathalyzer that is calibrated to be too sensitive.

7) Super Duper High School has a total of 143 teachers. Suppose that you are a researcher who is interested in studying Teacher Effectiveness at SDHS. You intend to evaluate the effectiveness of **all** of the teachers for your report.

a) What type of data collection method is this?

b) Suggest at least two **valid** variables that you might study. Include an instrument that can be used to measure your variables and the units.

c) Suggest at least two **invalid** variables that you might study. Include an instrument that can be used to measure your variables and the units.

8) For each of the following scenarios, determine whether the situation described is an experiment, observational study, census, or a sample survey.

a) The Super Spaz Energy drink company randomly selects 2% of the cans filled each day, and tests them for volume, ingredient content, and taste.

b) A government lobbyist analyzes the crime reports for the 4 counties in her community.

c) New advertisements are generally tried out on focus groups before investing a lot of money to pay for airtime on national TV.

d) Each student in Probability and Statistics will take the District Common Assessment as a final exam.

e) A teenager decides to evaluate how serious her parents are about her curfew by coming home 15 minutes late just to see what happens.

9) Pasquale's Big and Tall Shop sold 127 suits during the first quarter of this year, and 17 were returned. Marco's XXL Shop sold 268 suits during the same time period, and 27 were returned.

a) What were the number of returns for each shop? Which shop had a higher number of returns?

b) What were the rates of returns for each shop? Which shop had a higher rate of returns?

c) Which of these statistics gives a more clear representation of customer satisfaction? Explain.

Review Exercises

$$\text{Rate of Change} = \frac{\text{Amount of Change}}{\text{Original Amount}}$$

- *The original amount is not always the largest or the smallest amount.*
- *Multiply by 100, to change to a percent.*
- *Determine whether this change is an increase or a decrease.*

10) Jolene makes $12.45 per hour at her job. Last year she made $10.85. What percent of a raise did Jolene receive?

11) Michaela's favorite shoes are normally $42.99. Today she found a sale in which they were marked down to $27.99. What percent of a discount is this?

12) The number of incidents of hazing reported at Some Random High School was 84 during the 2010-2011 school year. The following year there were 37 incidents of hazing reported at SRHS. What is the rate of change in reported hazing incidents between these two school years? Is it an increase or a decrease?

13) SRHS has had a huge problem getting students to class on time, so the administrators have implemented a new tardy policy. In an effort to determine whether or not it is working to deter students from being tardy to class, data has been collected and analyzed. The following table shows some of the data:

School Year	2010-2011	2011-2012	% of change (+ or –)
Total number of tardies (to any class period)	5186	4295	
Number of students with more than 10 tardies	175	59	
Number of students with more than 20 tardies	112	77	

a) Calculate the percent of change for each category and complete the table (round to the nearest tenth of a percent).

b) Which category saw the most significant change?

c) Based on these calculations, do you feel that the tardy policy is working? Explain your reasoning.

4.2 Sample Survey and Census

Learning Objectives

- Differentiate between population and sample
- Understand the terminology of sampling methods
- Identify various sampling methods
- Recognize and name sources of bias or errors in sampling



Population vs. Sample

What is the approval rate of the President? If we really wanted to know the true approval rating of the president, we would have to ask every single adult in the United States her or his opinion. If a researcher wants to know the exact answer in regard to some question about a population, the only way to do this is to conduct a census. In a **census**, every unit in the population being studied is measured or surveyed. In this example our **population,** the entire group of individuals that we are interested in, is every adult in the United States of America.

A census like this (asking the opinion of every single adult in the United States) would be impractical, if not impossible. First, it would be extremely expensive for the polling organization. They would need a large workforce to try and collect the opinions of every single adult in the United States. Once the data is collected, it would take many workers many hours to organize, interpret, and display this information. Other practical problems that might arise are: some adults may be difficult to locate, some may refuse to answer the questions or not answer truthfully, some people may turn 18 before the results are published, others may pass away before the results are published, or an event may happen that changes peoples' opinions drastically, etc. Even if this all could be done within several months, it is highly likely that peoples' opinions will have changed. So by the time the results are published, they are already obsolete.

Another reason why a census is not always practical is because a census has the potential to be destructive to the population being studied. For example, it would not be a good idea for a biologist to find the number of fish in a lake by draining the lake and counting them all. Also, many manufacturing companies

test their products for quality control. A padlock manufacturer, for example, might use a machine to see how much force it can apply to the lock before it breaks. If they did this with every lock, they would have none to sell. In both of these examples it would make much more sense to simply test or check a sample of the fish or locks. The researchers hope that the sample that they select represents the entire population of fish or locks.

This is why sampling is often used. **Sampling** refers to asking, testing, or checking a smaller sub-group of the population. A **sample** is a representative subset of the population, whereas the population is every single member of the group of interest. The purpose of a sample is to be able to generalize the findings to the entire population of interest. Rather than do an entire census, samples are generally more practical. Samples can be more convenient, efficient and cost less in money, labor and time.

A number that describes a sample is a **statistic**, while a number that describes an entire population is a **parameter**. Researchers are trying to approximate parameters based on statistics that they calculate from the data that they have collected from samples. However, results from samples cannot always be trusted.

Example 1



A poll was done to determine how much time the students at SDHS spend getting ready for school each morning. One question asked, "Do you spend more or less than 20 minutes styling your hair for school each morning?" Of the 263 students surveyed, 61 said that they spend more than 20 minutes styling their hair before school. Identify the population, the parameter, the sample, and the statistic for this specific question.

Solution

a) population (of interest): all students at SDHS

b) parameter (of interest): what proportion of students spend more than 20 minutes styling their hair for school each morning

c) sample: the 263 SDHS students who were surveyed

d) statistic: $\hat{p} = 61/263 = 0.2319 = 23.19\%$

Randomization

One common problem in sampling is that the sample chosen may not represent the entire population. In such cases, the statistics found from these samples will not accurately approximate the parameters that the researchers are seeking. Samples that do not represent the population are **biased**. If someone was interested in the average height of all male students at his or her high school, but somehow the sample of students measured included the majority of the varsity basketball team, the results would certainly be biased. In other words, the statistics that were calculated would most certainly overestimate the average height of male students at the school. Samples should be selected **randomly** in order to limit bias. Also, if only three students' heights are measured, it is very possible that the average height of these three will not be close to the average height of all of the male students. The average of the heights of 40 randomly

chosen male students would be more likely to result in a number that will match the average of the entire population than that of just three students. Larger sample sizes will have less variability, so small sample sizes should be avoided.

A **random sample** is one in which every member of the population has an equal chance of being selected. There are many ways to make such random selections. The way many raffles are done is that every ticket is put into a hat (or box), then they are shaken or stirred up, and finally someone reaches into the hat without looking and selects the winning ticket(s). Flipping a coin to decide which group someone belongs in is another way to choose randomly. Computers and calculators can be used to make random selections as well. The purpose of choosing randomly is to avoid any personal bias from influencing the selection process. Randomization will limit bias by mixing up any other factors that might be present. Think of the heights of those male students, if we assigned every male at that school a number and then had a computer program select 40 numbers at random, it is most likely that we would end up with a mixture of students of various heights (rather than a bunch of basketball players). Also, no one did something like just measuring their friends heights, or the first 40 males he sees who are staying after school, or everyone in first lunch who is willing to come be measured. A computer program has no personal stake in the outcome and is not limited by its comfort level or laziness.

If the goal of our sample is to truly estimate the population parameter, then some planning should be done as to how the sample will be selected. First of all, the list of the population should actually include every member of the population. This list of the population is called the **sampling frame**. For example, if the population is supposed to be all adults in a given city and someone is working from the phone book to make selections, then everyone who is unlisted and those who do not have a land line telephone will not have any chance of being selected. Therefore, this is not an accurate sampling frame.

Good Sampling Methods

Simple Random Sample

When the selection of which individuals to sample is made randomly from one big list, it is called a **simple random sample** (or SRS). An example of this would be if a teacher put every single student's name in a hat and then draws 5 names from the hat, without looking, to receive a piece of candy. In an SRS every single member of the population has an equal probability of being selected - every student has an equal chance of getting the candy. And, in an SRS every combination of individuals also has an equal chance of being selected - any group of 5 students might end up getting candy. It might be all 5 girls, it might be the 5 students who sit in the back row, or it might even end up being the 5 students who misbehave the most. Anything is possible with an SRS!

Stratified Random Sample

A simple random sample is not always the best choice though. Suppose you were interested in students' opinions regarding the homecoming theme, and you wanted to make certain that you heard from students from all four grades. In such a case it would make more sense to have four separate lists (freshmen, sophomores, juniors and seniors), and then to randomly select 50 students from each list to give your survey to. A selection done in this way is called a stratified random sample. A **stratified random sample** is when the population is divided into deliberate groups called **strata** first, then individual SRS's are selected from each of the strata. This is a great method when the researchers want to be sure to include data from specific groups. Divisions may be done by gender, age groups, races, geographic location, income levels, etc. With stratified random samples, every member of the population has an equal chance of being selected, but not every combination of individuals is possible.

Systematic Random Sample

Another way to choose a sample is systematically. A **systematic random sample** makes the first selection randomly and then uses some type of 'system' to make the remaining selections. A system could be: every 15th customer will be given a survey, or every 30 minutes a quality control test will be run. A systematic random sample might start with a single list like an SRS, randomly choose one person from the list, then every 25th person after that first person will also be selected. Systematic random samples still give every member of the population an equal chance of being chosen, but do not allow for all combinations of individuals. Some groups are impossible, such as a group including several people who are in order on the list.

Multi-Stage Random Sample

When seeking the opinions of a large population, such as all registered voters in the United States, a multi-stage random sample is often employed. A **multi-stage random sample** involves more than one stage of random selection and does not choose individuals until the last step. A pollster might start by randomly choosing 10 states from a list of the 50 states in the U.S.A. Then she might randomly choose 10 counties in each of those states. And, finally she can randomly choose 50 registered voters from each of those counties to interview over the telephone. When she is done, she will have 10x10x50 = 5000 individuals in her sample. This is another sampling method that gives individuals an equal chance of being chosen, but does not allow for all possible combinations of individuals. For example, there is no possible way that all 5000 of these voters will be from Texas.

Random Cluster Sample

Sometimes cluster samples are used to collect data. Splitting the population into representative **clusters,** and then randomly selecting some of those clusters, can be more practical than making only individual selections. In cluster sampling, a census is done on each cluster or group selected. When appropriately used, cluster sampling can be very useful and efficient. One needs to be careful that the clusters are in fact selected randomly and that this method is the best choice. When a study of teenagers across the country is to be done, a random cluster method can be the best choice. An SRS of all teens would be nearly impossible. Imagine that one big list of all teens! A multi-stage random sample might be theoretically ideal, but the practicality of surveying one teenager from a high school in Little Rock, and one from another high school in Duluth, and so on would be quite a nightmare. The best choice might be to randomly select 10 metropolitan areas, 10 suburban areas, and 10 urban areas from across the country. And then to randomly select one high school in each of these areas and then finally to randomly select 4 second hour classes from each of those high schools. Then survey the entire classes selected (clusters). This would be a combination of mulit-stage random selection and cluster sampling. Another use for random cluster sampling is quality control at a popcorn factory. If every hour, a bucket of popcorn is scooped out. The entire bucket of popcorn can be checked for salt content, appearance, number of kernels not popped, not burnt, etc. This is an example of a systematic random cluster sample, the system being 'every hour' a sample is taken, and the clusters being each bucket of popcorn.



Bad Sampling Methods



Voluntary Response Sample

Beware of call in surveys, and online surveys! Suppose that a radio hosts on KDWB says something like, "*Do you think texting while driving should be illegal? Call in and have your opinion heard!*" It is highly likely that many people will call in and vote "*No!*" However, the people who do take the time to call will not represent the entire population of the twin cities and so the results cannot possibly be trusted to be equal to what all members of the population think. The 'statistic' that this 'survey' calculates will be biased. The only people who will take the time to call in are those who feel strongly that texting while driving should be legal (or illegal). Such a sampling method is called a voluntary response sample. In **voluntary response samples**, participants get to choose whether or not to participate in the survey. Online, text-in, call-in, mail-in, and surveys that are handed out to people with an announcement of where to turn them in when completed, are all examples of voluntary response surveys. Voluntary response samples are almost always biased because they result in no response whatsoever from most people who are invited to complete the survey. So, most opinions are never even heard, except for those who have really strong opinions for or against the topic in question. Also, those who have strong opinions can call or text multiple times. A new problem that comes with the Internet is that many companies are offering to pay people to complete surveys, which makes any results suspect. For these reasons, the results of voluntary response samples are always suspect because of the potential for bias.

Convenience Sample

Another commonly used, but dangerous method for choosing a sample is to use a convenience sample. A **convenience sample** just asks those individuals who are easy to ask or are conveniently located - right by the pollster for example. The big problem here is that the sample is unlikely to represent the entire population. The fact that this group was convenient, implies that they most likely have at least something in common. This will almost always result in biased results. An interviewer at the mall only asks people who shop at the mall, and only at some given time of day, so many people in the community will never have the opportunity to be interviewed. When the population of interest is only mall-shoppers, this will be somewhat better than when the population of interest is community members. Even then, the interviewers choose whom to go up to and the interviewees can easily refuse to participate.

With both of the bad sampling methods, the word random is nowhere to be found. That lack of randomness should serve as a big hint that some type of bias will likely be present. The scary thing is that most of the results we see published in the media are the results of convenience samples and voluntary response samples. One should always ask questions about where and how the data was collected before believing the reported statistics.

Example 2

Suppose that a survey is to be conducted at the new Twin's Stadium. A five question survey if developed. Population of interest: All of the 31,045 fans present that day. Sample size: 2,500 randomly selected fans. Identify specifically the sampling method that is being proposed in each scenario. Also, comment on any potential problem or bias that will likely occur.

a) The first 2,500 fans to arrive are asked five questions.

b) Fifty sections are randomly selected. Then ten rows are randomly selected from each of those sections. Then five seats are randomly selected from each of those rows. The people in these seats are interviewed in person during the game.

c) A computer program selects 2,500 seat numbers randomly from a list of all seats occupied that day. The people in these seats are interviewed in person during the game.

d) 2,500 seats are randomly selected. The surveys are taped to those 2,500 with instructions as to where to return the completed surveys.

e) The number 8 was randomly selected earlier. The 8th person through any gate is asked five questions. Then, every 12th person after that is also asked the five questions.

f) The seats are divided into 25 sections based on price and view. A computer program randomly selects 100 seats from each of these sections. The people in these seats are interviewed in person during the game.

Solution

a) This is a convenience sample. It will not represent everyone present that day. This will suffer from bias because all of these people have at least one thing in common- they arrived early.

b) This is a multi-stage random sample. It will probably represent the entire population. As long as the people are in their seats and willing to answer the questions honestly, it could be a good plan.

c) This is simple random sample. It will probably represent the entire population. As long as the people are in their seats and willing to answer the questions honestly, it could be a good plan.

d) This is a voluntary response sample. It is very likely that most of those surveys will end up on the ground or in the garbage. This will likely suffer from many people not responding. It is also probable that anyone who had an extremely negative experience will be more likely to complete their surveys.

e) This is a systematic random sample. It will probably represent the entire population. As long as the people are willing to answer the questions honestly, it could be a good plan.

f) This is a stratified random sample. It will probably represent the entire population. As long as the people are in their seats and willing to answer the questions honestly, it could be a good plan.

Errors in Sampling

Sampling Errors

Some errors have to do with the way in which the sample was chosen. The most obvious is that many reports result from a **bad sampling method**. Convenience samples and voluntary response samples are used often and the results are displayed in the media constantly. Now, we have seen that both of these methods for choosing a sample are prone to bias. Another potential problem is when results are based on **too small of a sample**. If a statistic reports that 80% of doctors surveyed say something, but only five doctors were even surveyed this does not give us a good idea of what all doctors would say.

Another common mistake in sampling is to leave an entire group (or groups) out of the sample. This is called **undercoverage**. Suppose a survey is to be conducted at your school to find out what types of music to play at the next school dance. The dance committee develops a quick questionnaire and distributes it to 12 randomly selected 5th period classes. However, what if they did this on a day when the football teams and cheerleaders had all left early to go to an out of town game. The results of the dance committee's survey will suffer from undercoverage, and will therefore not represent the entire population of your school.

There is also the fact that each sample, randomly selected or not, will result in a different group of individuals. Thus, each sample will end up with different statistics. This expected variation is called **random sampling error** and is usually only a slight difference. However, every now and then the sample selected can be a 'fluke' and just simply not represent the entire population. A randomly selected sample might accidentally end up with way too many males for example. Or a survey to determine the average GPA of students at your school might accidentally include mostly honor's students. There is no way to avoid random sampling error. This is one reason that many important surveys are repeated with a new sample. The odds of getting such a 'fluke' group more than once are very low.

Non Sampling Errors

One of the biggest problems in polling is that most people just don't want to bother taking the time to respond to a poll of any kind. They hang up on a telephone survey, put a mail-in survey in the recycling bin, or walk quickly past an interviewer on the street. Even when the researchers take the time to use an appropriate and well-planned sampling method, many of the surveys are not completed. This is called **non-response** and is a source of bias. We just don't know how much the beliefs and opinions of those who did complete the survey actually reflect those of the general population, and, therefore, almost all surveys could be prone to non-response bias. When determining how much merit to give to the results of a survey, it is important to look for the response rate $\left(\frac{number returned}{total number sent}\right)$.

The wording of the questions can also be a problem. The way a question is worded can influence the response of those people being asked. For example, asking a question with only two answer choices forces a person to choose one of them, even if neither choice describes his or her true belief. When you ask people to choose between two options, the order in which you list the choices may influence their response. Also, it is possible to ask questions in leading ways that influence the responses. A question can be asked in different ways which may appear to be asking the same thing, but actually lead individuals with the same basic opinions to respond differently. Consider the following two questions about gun control.

> "*Do you believe that it is reasonable for the government to impose some limits on purchases of certain types of weapons in an effort to reduce gun violence in urban areas?*"

> "*Do you believe that it is reasonable for the government to infringe on an individual's constitutional right to bear arms?*"

The first question will result in a higher rate of agreement because of the wording 'some limits' as opposed to 'infringe'. Also, 'an effort to reduce gun violence' rather than 'infringe on an individual's constitutional right' will bring more agreement. Thus, even though the questions are intended to research the same topic, the second question will render a higher rate of people saying that they disagree. Any person who has strong beliefs either for or against government regulation of gun ownership will most likely answer both questions the same way. However, individuals with a more tempered, middle position on the issue might believe in an individual's right to own a gun under some circumstances, while still feeling that there is a need for regulation. These individuals would most likely answer these two questions differently.

You can see how easy it would be to manipulate the **wording of a question** to obtain a certain response to a poll question. This type of bias may be done intentionally in an effort to sway the results. But it is not necessarily always a deliberate action. Sometimes a question is poorly worded, confusing, or just plain hard to understand, this will still lead to non-representative results. another thing to look at when critiquing the results of a survey is the specific wording of the questions. It is also important to know who paid for or who is reporting the results. Do the sponsors of this survey have an agenda they are trying to push through?

A major problem with surveys is that you can never be sure that the person is actually responding truthfully. When an individual responds to a survey with an incorrect or untruthful answer, this is called **response bias** . This can occur when asking questions about extremely sensitive, controversial or personal issues. Some responses are actual lies, but it is also common for people just to not remember correctly. Also, sometimes someone who is completing a survey or answering interview questions will 'mess with the data' by lying or making up ridiculous answers.

Response bias is also common when asking people to remember what they watched on TV last week, or how often they ate at a restaurant last month, or anything from the past. Someone may have the best intentions as they complete the questionnaire, but it is very easy to forget what you did last week, last

month, or even yesterday. Also, people are often hurrying through survey questions, which can lead to incorrect responses. So the results on questions regarding the past should be viewed with caution.

It is difficult to know whether or not response bias is present. We can look at how questions were worded, how they were asked, and who asked them. Person-to-person interviews on controversial topics carry a definite potential for response bias for example. It is sometimes helpful to see the actual questionnaire that the subjects were asked to complete.

There are sometimes mistakes in calculations or typos present in results, these are **processing errors** (or human errors). For example, it is not uncommon for someone to enter a number incorrectly when working with large amounts of data, or to misplace a decimal point. These types of mistakes happen frequently in life, and are not always caught by those responsible for editing. If a reported statistic just doesn't seem right, then it is a good idea to recheck calculations when possible. Also, if the numbers appear to be 'too good to be true', then they just might be!



Example 3

The department of health often studies the use of tobacco among teens. The following is a description by the Minnesota Department of Health describing how they chose the sample for the 2008 Minnesota Youth Tobacco and Asthma Survey. In 2008, they had 2,267 high school students complete surveys and 2,322 middle school students complete surveys. Each student in the sample completed an extensive questionnaire consisting of many questions related to tobacco use. Answer the questions that follow. To see the entire report go to: http://www.health.state.mn.us/divs/hpcd/tpc/reports/

> Students were selected for the survey in two stages. First, 48 public middle schools (grades 6-8) and 51 public high schools (grades 9-12) were randomly selected, with probability of selection based on size of enrollment. Alternative schools and charter schools were included. The sample schools were randomly chosen by CDC using enrollment data provided by the Minnesota Department of Health. Next, three or four classrooms within each participating school were randomly selected, and all students in these classrooms were invited to participate. The number of schools and classrooms selected was reduced substantially in 2008 in order to reduce the burden on schools. The sample size is still adequate to provide reasonable statewide estimates.

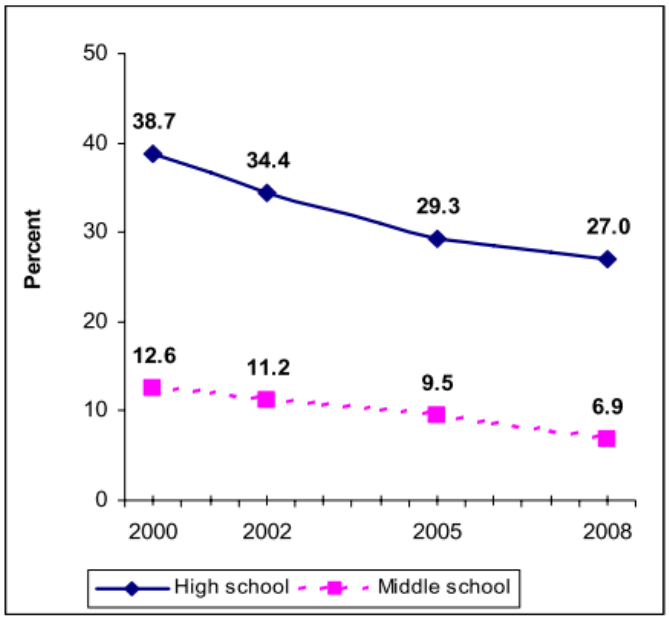
a) What type of sampling methods were used for this?

b) Identify the population, the parameter of interest, the sample, and the individuals for this study

c) When asking teens about tobacco use, what types or causes of bias will likely be present?. What could be done to limit bias?

d) This graph shows how the percent of teens using tobacco has changed from 2000 to 2008. Identify the statistics that were found for this question in 2008.

Percent using any tobacco in last 30 days



Solution

a) This study used a complicated combination of sampling methods. They used a stratified, multi-stage, random cluster sample method to select individuals. It was a stratified random sample (by high school and middle school), it was a multi-stage random sample (first random schools were selected, second random classes were selected), and it was a cluster sample (every student in each class was given the survey).

b) population (of interest): all middle and high school students in Minnesota

parameter (of interest): teen tobacco use

sample: 2267 high school, and 2322 middle school students in Minnesota (from 48 public middle schools and 51 public high schools in Minnesota)

individuals: each student who completed a survey

c) Response bias: Tobacco use is not legal for people under 18, so teens will not want to tell the truth if they think they may get in trouble.

Non-response bias: Some people were absent the day survey was given.

Undercoverage: Only public school students were included, so those who attend private schools were left out.

Wording of the questions: This could be a problem, but we do not know the exact wording so cannot be sure.

To avoid the response bias factor, surveys regarding controversial topics should all be anonymous. If you read further into this report, you will see that the students were assured all results would be anonymous (no names or ID numbers included).

To avoid the non-response bias factor, students who were absent could be given the survey when they return to school.

To avoid the undercoverage of private school students, private schools could be included in the sample.

d) In 2008, 27.0% of the high school students asked and 6.9% of the middle school students asked had used tobacco in the last 30 days.

Problem Set 4.2

Section 4.2 Exercises

1) For each of the following, determine whether the bold number is a ***parameter*** or a ***statistic.*** *(hint: remember that a parameter is a number that represents an entire population and a statistic is a number that represents a sample)*

a) The average height of all oak trees is **42.3 feet**.

b) Ms. Anderson's class average on the final exam was **71.4%**.

c) The average number of songs that the students surveyed have on their iPods was **791 songs**.

d) Itunes reports that the average number of songs people have on their iPods is **503 songs**.

e) The sticker on the Super Speedster Sport Sedan says **17.82 mpg**.

f) Martin had to keep track of how much time he spent watching TV for a whole week. He found that last week he averaged **3.4 hours of TV per day**.

2) Minnesota's Best High School found that last year they did not have enough seats or room for all of the family members who wished to attend the graduation ceremony. The administrators at MBHS need to decide where to hold the graduation ceremony this year, so they sent a questionnaire home with each of this year's 543 seniors early in September. They asked for the surveys to be completed and returned by September 27th. Of the 148 surveys returned, the average number of seats that will be needed is 6.2. To be safe, the administrators use 7 and determine that they will need a hall that can hold 3800 people (543 students X 7 seats = 3801 seats needed). Using this number they find an appropriately sized hall and reserve it. Identify each of the following as specifically as possible.

a) population (of interest)

b) parameter (of interest)

c) sample

d) statistic

e) sampling method that was used

f) What is the response rate (the percent of surveys returned)?

g) What is wrong with the what these administrators have done? What type of bias or error is likely present?

h) Will the statistic most likely by too high or too low? What is a likely consequence of this biased result?

3. Suppose that a survey is to be conducted at Minnesota's Best High School. Population of interest: 2640 MBHS students. Sample size: 240 MBHS students. Identify specifically the sampling method that is being proposed in each scenario. Also, comment on any potential problem or bias that will likely occur.

a) Every freshman's name is put on a slip of paper and put into a giant bucket. Sixty names are pulled out of the hat. This process is repeated for each grade level.

b) A list of all students is obtained from the counselors. Julie randomly selects a number between 1 and 2640 and then finds the student that matches this number on the list. She then selects every eleventh person on the list after that one (cycling back to the beginning of the list) until 240 names are chosen.

c) Surveys are handed out with lunches. Students are asked to complete them and turn them in on a table in the front of the cafeteria.

d) A computer randomly selects 240 names from the entire list of students in the school database.

e) Twelve teachers are randomly selected. Two of each of their classes are then randomly selected. Ten students from each of these classes are then selected.

f) Three teachers, Mr. Niceguy, Mr. Greatguy and Mr. Happyguy, each volunteers to survey the students in all of his classes.

4. Name, and briefly describe, the type of bias that would most likely be present in each of the following situations:


I'M FILLING OUT A READER SURVEY FOR CHEWING MAGAZINE.
SEE, THEY ASKED HOW MUCH MONEY I SPEND ON GUM EACH WEEK, SO I WROTE, "$500." FOR MY AGE, I PUT "43," AND WHEN THEY ASKED WHAT MY FAVORITE FLAVOR IS, I WROTE "GARLIC / CURRY."
THIS MAGAZINE SHOULD HAVE SOME AMUSING ADS SOON.
I LOVE MESSING WITH DATA.

a) What is the name of the type of bias in the cartoon?

b) As the 2010 Census was being conducted, many people did not return their forms. What type of bias is this?

c) What type of bias would most likely be present if high school students are interviewed about their drinking and drug use habits? Would the statistics most likely over- or under-estimate the true parameters?

d) What is the one type of sampling error that we expect to happen, but cannot do anything to avoid, called?

e) When calculating the statistics from a survey, a typo is made. What type of error is this?

f) A radio talk show host asks, "Do you think that the driving age should be changed to 18?" What type of bias will most likely be present? Why is this?

g) If a survey is conducted by door-to-door interviews and the interviewers skip a few neighborhoods that 'make them nervous', what type of bias is this called?

h) If an interviewer asks each person, "*Do you prefer Pizza Ickarooni, or the delicious fresh flavors of Pizza Delicioso?*", what type of bias is present?

Review Exercises

5) One die is rolled, what is the chance that a number greater than four or an even number is showing?

6) One die is rolled, what is the chance that a number greater than four and an even number is showing?

7) Two dice are rolled. What is the probability that the sum of the number of dots showing is nine or greater?

8) If three dice are rolled, what is the probability of getting three of a kind (all 3 dice show the same number of dots)?

4.3 Random Selection



Learning Objectives

- Obtain a random sample using a random digit table
- Describe the process followed to obtain an SRS
- Outline an appropriate sampling method

Random Selection

We have discussed that it is important to choose samples randomly in order to reduce bias, but we haven't discussed how to actually carry out the process. There are many ways to make random selections. A common way to choose things at random is to use a 'big hat', or box, or bowl, etc. For example, suppose that a teacher wanted to randomly select 5 students every day, from a class of 34 students, to hand in their homework to be graded. Each day she has all of the students' names in a big fish bowl. She will mix the names up and select 5 names. These students will turn their homework papers in right then, and the other students will not need to. The five selected names will be put back in the fishbowl, so they may be selected again tomorrow. This is an example of an SRS of size 5 of her class. Every student has an equal probability (5/34 or 14.7% chance) of being required to turn in his or her homework on any given day and any combination of five students may be chosen. One student may end up turning in her assignment several days in a row, while another student may never need to turn hers in all year long. The idea of a 'big hat' is a good method for random selection when working with small populations, but it is not always practical.

Random selections can be made by flipping coins, rolling dice, or spinning a spinner. These days, most random selections can be done using technology such as a computer program or a **random number generator** on a graphing calculator, Another way that random selections are made in statistics is by using a random digit table. A **random digit table** is a long list of randomly generated digits from 0 to 9. The digits are listed in groups of five simply to make it easier to read and not lose your place. Imagine that someone has a ten-sided die with each digit from 0 to 9 marked on a side. They sit down, roll the die and write down the digit that appears, then they roll it again and write down the digit that appears, then they do this again and again. As you can imagine, this would take quite awhile, but would result in a long list of random digits. This is basically what a random digit table is.

How to Use a Table of Random Digits

There is a process to follow when using a random digit table to make your selection. You need to report your process with enough detail that if someone else were to follow your steps, they would end up with the exact same randomly selected numbers. The purpose of this is to prove, if needed, that your selection process was truly random so that no one can accuse you otherwise. The following example illustrates the steps you will need to follow (and explain) when using a random digit table to make your random selection.

Example 1

Five boxes, each containing 24 cartons of strawberries, are delivered in a shipment to a grocery store. The produce manager always selects a few cartons randomly to inspect. He knows better than to just look at some of the cartons on the top or only in one box, because sometimes the rotten ones are on the bottom. Today he wishes to select a total of 6 cartons to inspect. He has the boxes arranged in order and has a set way to count the cartons inside each box. Explain the process used to make the random selection using a random digit table.

Solution

Step 1: Assign numbers to the list (must all be an equal number of digits long)

Since he has 120 cartons total, he will assign the numbers 001 to 120 to represent the cartons in order.

Step 2: Choose a starting line on the random digit table. If the problem states a line to start at, use that line. Otherwise, pick any line you want and record the line number. If you run out of digits, simply move to the next line down.

He will use line 119 to make the selections.

Step 3: Decide how many digits to look at each time. The number of digits in your largest number is required.

He will need to look at 3-digit numbers every time.

Step 4: Decide if any numbers will need to be ignored and whether or not repeats will be allowed.

He will not want to inspect the same carton twice, so he will ignore any repeats. And, any numbers above 120 will not apply in this case, so he will ignore numbers 121-999 and 000.

Step 5: State when to stop.

He will stop once six numbers are selected. He will then find the cartons that the numbers represent and inspect those cartons.

Step 6: Report the numbers that were selected. When given a specific list, go back and determine which specific individuals have been selected.

Here is a part of the random digit table so that you can see how the selection was made. Note that dividers have been placed between each group of 3-digits for this example. When we reach the end of a line, we simply continue on the following line.

Line #	*random digits in groups of five:*								*selections:*
Line 119	958|57	0|711|8	87|664|	920|99	5|880|6	66|979|	986|24	8|482|6	no values fit the range
Line 120	35|476|	559|72	3|942|1	65|850|	**042**|66	3|543|5	43|742|	**119**|37	#042 and #119
Line 121	7|148|7	09|984|	290|77	1|486|3	61|683|	470|52	6|222|4	51|**025**|	#025
Line 122	138|73	8|159|8	95|**052**|	909|**08**	**7**|359|2	75|186|	871|36	9|576|1	#052 and #087
Line 123	54|580|	815|**07**	**2**|7102	56027	55892	33063	41842	81868	#072 that is six, so we stop

So, the strawberries in cartons numbered 042, 119, 025, 052, 087, and 072 will be inspected. The entire deliverywill be accepted or rejected based on this random sample of 6 cartons.



Example 2

Five of the employees at the Stellar Boutique are going to be selected to go to a training in Las Vegas for four days. Everyone wants to go of course, so the owner has decided to make the selection randomly. She has decided to send two managers and three sales representatives. The employees' names are listed in the table below.

a) What type of sampling method is this?

b) Explain the process she can follow to use a random digit table, starting at line #108, to select the employees who will get to go to the training. Select the managers first, then select the sales representatives.

Managers
Angela
Barbara
Elise
Gigi
Malena
Rosie
Tammy
Veronica

Sales Representatives	**Sales Representatives**	**Sales Representatives**
Alfie	Ilma	Ray Anne
Bettie Lou	Jo Jo	Sandy
Cari	Katarina	Shirley
Carry	Lin	Suzi
Darcy	Marcie	Tawanda
Fan Fan	Nancy	Wendy
Heidi	Oprah	Zulu

Solution

a) This is a stratified random sample.

b) **For the managers:**

- ***Assign numbers to the list 1 to 8***
- ***Use random digit table, starting at line #108***
- ***Look at one digit at a time***
- ***Ignore 9, 0, and any repeats***
- ***Stop when two have been selected***
- ***State the names***

Managers
1- Angela
2- Barbara
3- Elise
4- Gigi
5- Malena
6- Rosie
7- Tammy
8- Veronica

Line #	*Random digits in groups of five*	*Selections*
Line 108	**6**|0|9|**4**|0 72024 17868 24943 61790 90656 87964 18883	#6 and #4

So, Rosie(#6) and Gigi(#4) will be the managers who get to go to Las Vegas.

For the sales representatives:

- ***Assign numbers to the list 01-21***
- ***Use random digit table, starting on the next line, #109***
- ***Look at two digits at a time***
- ***Ignore 22-99, 00, and any repeats***
- ***Stop when three have been selected***
- ***State the names***

Sales Representatives	**Sales Representatives**	**Sales Representatives**
01-Alfie	08- Ilma	15- Ray Anne
02- Bettie Lou	09- Jo Jo	16- Sandy
03- Cari	10- Katarina	17- Shirley
04- Carry	11- Lin	18- Suzi
05- Darcy	12- Marcie	19- Tawanda
06- Fan Fan	13- Nancy	20- Wendy
07-Heidi	14- Oprah	21- Zulu

Line #	*Random digits in groups of five*	*Selections*
Line 109	36|00|9 1|93|65| **15**|41|2 3|96|38| 85|45|3 4|68|**16**| 83|48|5 4|**19**|79	#15, #16, & #19

So, Ray Anne(#15), Sandy(#16), and Tawanda(#19) will be the sales representatives who get to go to Las Vegas.

Problem Set 4.3

Section 4.3 Exercises

Use the table of random digits in the appendix for the following problems.

1. The manager at Big-N-Nummy-Burger wishes to know his employees' opinions regarding the work environment. He has 56 employees and plans to select 12 employees at random to complete a survey.

a) Explain the process he can follow to use a random digit table, starting at line 108, to select an SRS of size 12.

b) Which employees numbers were selected?

2. Use a random digit table to select an SRS of five of the fifty U.S. States. Explain your process thoroughly and report the five states that you chose. Repeat this a second time, but begin on a different line on the random digit table. Compare your lists to another classmate's lists. Did you end up with any of the same states in your samples?

Table 4.1:

Alabama	Alaska	Arizona	Arkansas
California	Colorado	Connecticut	Delaware
Florida	Georgia	Hawaii	Idaho
Illinois	Indiana	Iowa	Kansas
Kentucky	Louisiana	Maine	Maryland
Massachusetts	Michigan	Minnesota	Mississippi
Missouri	Montana	Nebraska	Nevada
New Hampshire	New Jersey	New Mexico	New York
North Carolina	North Dakota	Ohio	Oklahoma
Oregon	Pennsylvania	Rhode Island	South Carolina
South Dakota	Tennessee	Texas	Utah
Vermont	Virginia	Washington	West Virginia
Wisconsin	Wyoming		

3. Washington High School has had some recent problems with students using steroids. The district decides that it will randomly test student athletes for steroids and other drugs. The boy's hockey team is to be tested. There are 13 players on the varsity team and 21 players on the junior varsity team. Use a table of random digits starting at line 122, to choose a stratified random sample of 3 varsity players and 5 junior varsity players to be tested. Remember to clearly describe your process.



Varsity Team Roster, by last name:	
Alexander	Nix
Baker	Radamacher
Brooks	Ritchie
Finch	Smithe
Gustaf	Thomas
Linder	West
Mullen	

Junior Varsity Team Roster, by last name:	
Andersen	Manzel
Anderson	Peterson
Baker	Randal A
Christian	Randal J
Donnovan	Reeder
Greene	Rice
Hansen	Sams
James	Sentel
Klein	Thorne
Linder	West
Lutz	

Review Exercises

4) Sketch a Venn Diagram that shows two events that are mutually exclusive.

5) Suppose that a survey was conducted at SRHS and it found that 86% of students have their own cell phones, and that 64% of students have their own IPod (or other similar personal music device). Furthermore, 9% of the students at SRHS say that they have neither one of these.

a) Define your variables and construct a Venn Diagram that fits this scenario.

b) What is the probability that a randomly selected student has both a cell phone and an IPod?

c) What is the probability that a randomly selected student has either a cell phone or an IPod?

4.4 Statistical Conclusions



Learning Objectives

- Understand when valid statistical conclusions can be made.
- Calculate an estimated margin of error and 95% confidence interval
- Make confidence statements

Statistical Conclusions

Remember that when you collect information from every unit in a population, it is called a census. In doing a census, we can be certain that the numbers we have calculated really do represent the entire population. But, because a census is often impractical, we generally take a representative sample of the population, and use that sample to try to make conclusions about the entire population. The downside to sampling is that we can never be completely, 100% sure that we have captured the truth about the entire population.

For example, imagine taking a random sample of 100 from a large population. Put those back and choose another sample of 100, repeating many times. Each of these samples of size 100 will include a different combination of 100 members of the population. Thus, each sample will result in different statistics. This natural difference between various samples is an expected random sampling error. To take this into account, researchers generally report their findings to have a **margin of error** or to be within a certain range of possible values. This range is called a **confidence interval.** For example the President's approval rating might be reported as, *"The approval rating for the President is 43.2%, with a margin of error of ±3%."* Which could also be reported as, *"The approval rating for the President is between 40.2% and 46.2%."*

Using a statistic to make a conclusion about a population is called statistical inference. This course is an introduction course, so we will only briefly touch on this idea. In a future statistics class, you will learn much more about statistical inference and calculations. It is important to note that statistical conclusions are meaningless when poor sampling techniques have been used. If the data was collected from a voluntary response sample, or you had a low response rate, or an incomplete sampling frame was used, then don't waste your time performing inference on your statistics. Random sampling error is the only type of error or bias that the margin of error accounts for.

95% Confidence Intervals

Once a statistic is calculated for a sample, it is used as an estimate for what the actual parameter might be. We do not know whether our statistic is close to the population parameter, or if it is too high, or too low, so we build our interval around the statistic. We add the margin of error to, and subtract the margin of error from, our statistic. We then report this range of values as our **confidence interval**, the interval that we are fairly confident that the true parameter must be within. In a more formal course you will learn how to calculate the margin of error more precisely, and for various levels of confidence (such as 90% or 99% etc.). In this course we will use a simple formula that estimates the margin of error for a

95% confidence interval. We will also make a **95% confidence statement**, which explains our conclusion regarding the population parameter in context. The formulas for an estimated 95% margin or error and confidence interval are:

Margin of error formula: $m.e. = \pm \frac{1}{\sqrt{n}}$

Confidence interval:

$\hat{p} \pm m.e.$ or statistic $\pm$ margin of error

n = *sample size*

**note: In order to make a smaller margin of error, and therefore a more narrow confidence interval, one must increase the size of the sample.*

Once you have found the range of numbers for your confidence interval, you are going to state your conclusion in context. Such a statement is called a **confidence statement**. The confidence interval refers to the population - not the sample. We are 100% certain of our sample statistic. It is the population parameter that we are estimating. Writing a confidence statement can be kind of confusing, so you can just use the following template:

> "**We are 95% confident that the true proportion of _____*(parameter of interest)*___-___ will be between _____*(low value of CI)*_____ and ___*(high value of CI)*___-_____."



Example 1

A random sample of 125 union members was conducted to see whether or not the union members would support a strike. Sixty-four of those surveyed said that they would support a strike unless safety conditions were improved. Identify each of the following as specifically as possible:

a) Population of Interest: All members of this union

b) Parameter of Interest: The percent of the union members who would support a strike

c) Sample: The 125 union members who were surveyed

d) Statistic: (p-hat) $= \frac{64}{125} = 0.512$

e) Margin of Error: m.e. $\pm\frac{1}{\sqrt{125}} = \pm 0.0894$

f) 95% Confidence Interval: 0.512 + 0.0894 = 0.6014 and 0.512 - 0.0894 = 0.4226

[0.4226 to 0.6014] or [42.26% to 60.14%]

g) Confidence Statement: ***"We are 95% confident that the true proportion of union members who would support a strike is between 42.26% and 60.14%"***



(Note- Two additional Videos in the Example)

Problem Set 4.4

Section 4.4 Exercises

1. A survey was done to determine the texting habits of MBHS students. An SRS of 270 students were asked several questions related to texting and cell phone usage. Of particular interest to the researchers was the proportion of students who text while in class. Of those surveyed, 178 said that they text during class at least ten times per week. Identify each of the following as specifically as possible.

a) Population of Interest

b) Parameter of Interest

c) Sample

d) Statistic

e) Margin of Error

f) 95% Confidence Interval

g) Confidence Statement

h) Do you personally feel that this is too high or too low of an estimate of the proportion of teens at your high school who text during class?

2. To predict the outcome of an upcoming Mayoral election, a random sample of 814 voters was selected. These people are asked several questions regarding the election. One question asked whether they were "..leaning Republican, Democratic, Independent, or other/undecided?" Based on this question, 38.2% of respondents said that they were "..leaning Democratic...". Identify each of the following as specifically as possible.

a) Population of Interest

b) Parameter of Interest

c) Sample

d) Statistic

e) Margin of Error

f) 95% Confidence Interval

g) Confidence Statement

3. In the same survey, 42.3% said that they were "...leaning Republican...".

a) Calculate an estimated 95% confidence interval

b) Is this enough evidence to "call" the election in favor of the republicans? Why or why not?

4. The quality control officer at Spaz Cola uses a systematic random sampling method to select cans of Spaz Cola to determine whether the machines are maintaining the correct recipe. Among the 480 cans analyzed today, 43 cans contained less sugar than the Spaz recipe requires! Identify each of the following as specifically as possible.

a) Population of Interest

b) Parameter of Interest

c) Sample

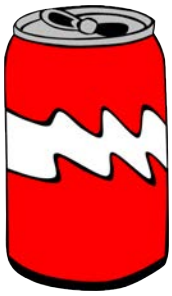
d) Statistic

e) Margin of Error

f) 95% Confidence Interval

g) Confidence Statement

h) Do you think that the company should be concerned? Why or why not?

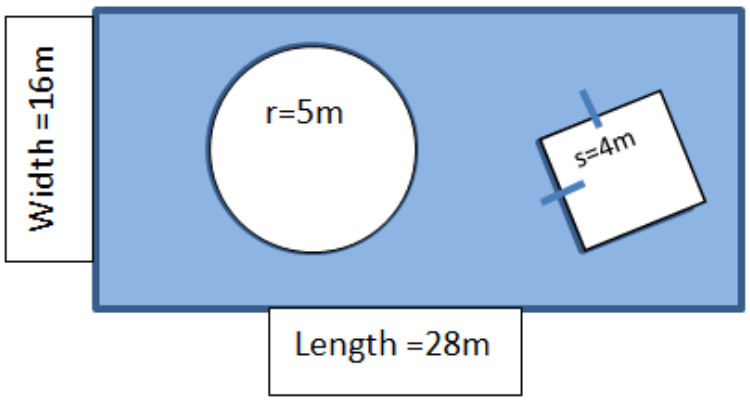


Review Exercises

5) Marcus got 18 points correct, out of 42 possible points, on his science test. On his history test, Marcus got 31 points out of 55 possible points. On which test did Marcus do better? Explain or show how you know.

6) Lydia got 15 points correct on her probability quiz (out of 23 possible). Then she earned 37 points, of the 48 possible points on her probability test. On which of these assessments did Lydia do better? Explain or show how you know.

7) The figure below is a dartboard. Suppose that a dart is thrown at it randomly. What is the probability that the dart will land on the shaded area?



8) Sketch two different ”dart boards” such that the probability of hitting the shaded are is equal to one-third.

4.5 Experiments and Observational Studies



Learning Objectives

- Know the terminology of basic experimental design
- Identify the elements of an experiment
- Distinguish between observational studies and experiments
- Outline experiments
- Understand the effects of lurking variables

Observational Studies and Experiments

When researchers collect data about subjects without imposing any type of treatment, they are doing an **observational study**. Many conclusions have been based on observational studies. The discovery that smoking causes lung cancer was initially theorized based on observational studies. Many consumers of cigarettes and tobacco companies questioned the validity of such studies, suggesting that it could have been some other variables that caused the cancers, not the cigarettes. **Retrospective studies**, based on past history of lung cancer patients showed that a high proportion of them were smokers. This did not convince those who either enjoyed smoking, or were making money off of tobacco. There could be some **lurking variables** to blame, extra variables that were not taken into account, but were actually the cause. **Prospective studies**, following people in the future, were undertaken in an effort to see whether or not there was a link between cigarette smoking and lung cancer. The statistics were still called into question because statisticians know that the only way to truly show causation is through a controlled experiment.

An **experiment** imposes some 'treatment' on the subjects. A **controlled** experiment involves having more than one group, where the only variable that is different between the groups is the treatment being tested. And, subjects will need to be assigned at **random** (left to impersonal chance) to the various treatment groups to control for lurking variables. With regard to cigarettes and lung cancer, researchers would need to find a group of non-smokers and randomly divide them into two groups. The randomization will divide up lurking variables that the researchers cannot control for. Also, there needs to be a fairly large number of subjects in each group so that the results do not appear to be some kind of a fluke. The researchers would then need to force one group to smoke cigarettes, while making sure that those in the control group did not smoke. This would go on for several years and both groups would need to be checked for lung cancer regularly. Clearly, there is no ethical way to do such an experiment. We cannot force people to do something that we suspect may cause cancer! Scientists were able to experiment on rats to see whether or not cigarettes caused cancer, and it did. Eventually, the compilation of all of these studies convinced everyone that smoking does cause cancer.

The Three Elements of Good Experimental Design are:

1. **Randomization**–Subjects must be randomly assigned to treatment groups in an effort to divide up any lurking variables

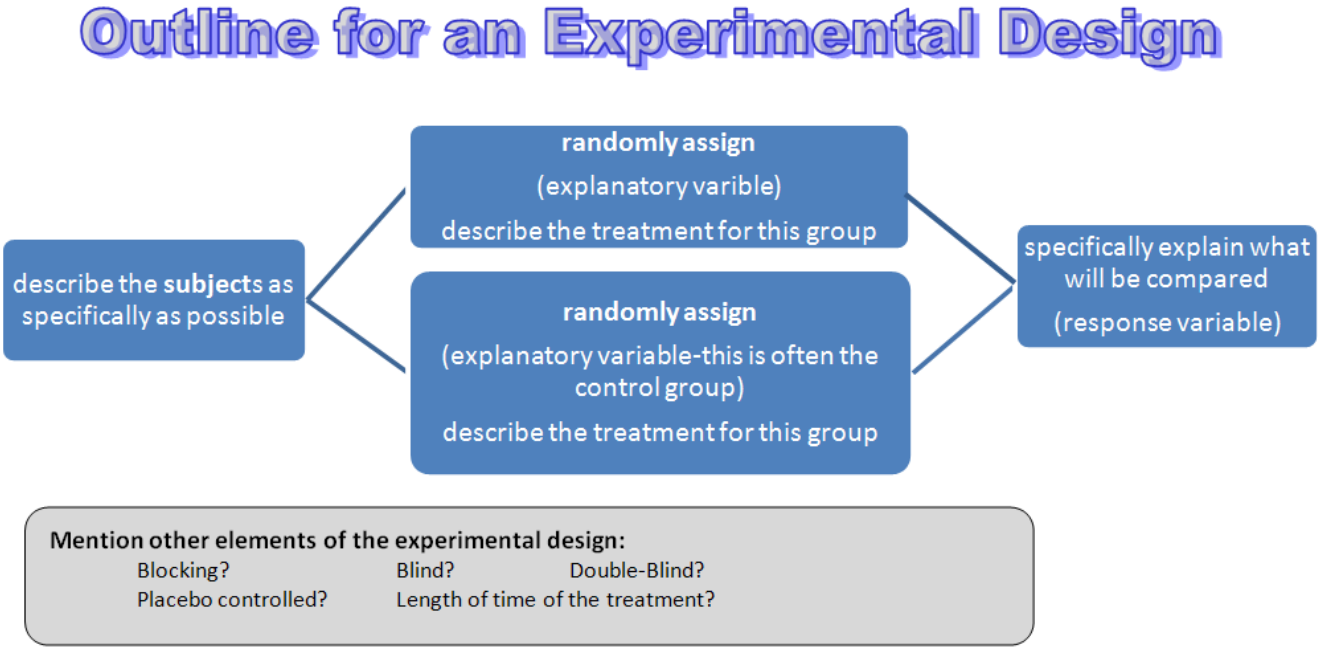
2. **Control**–There should be a control group– a group that does not receive the treatment. Having more than one group, where the only difference is the treatment being tested, allows for comparisons to be made.
3. **Replication**–There should be a large enough number of subjects so that the results seem believable. Also, the experiment should be able to be replicated on a different group of subjects.



Experimental Design

In an experiment, the people, animals, or objects, that are being experimented on are called the **subjects**. The treatment that is being tested is the **explanatory variable**. The result, outcome, or change that happens (or doesn't happen) is the **response variable**. Keep in mind that sometimes it is necessary to give a pre-test prior to imposing the treatment. For example, if we are testing a medication that claims to lower cholesterol levels, we will certainly need to know the cholesterol levels of all of our subjects prior to giving them the treatment. At the end of the experiment we will again test them and then we can compare any change in cholesterol level.

The control group may be given no treatment at all. Or, you may want to use the control group as a way to compare a new treatment to an old treatment. For example, if someone has developed a new medication that they believe will cure headaches, they will want to compare it to aspirin, acetaminophen, and ibuprofen. Such researchers will likely form four randomly assigned groups (Groups A, B, C, and D), assigning the subjects in each respective group to take a specific one of the treatments whenever they have a headache and to record whether or not it worked and how quickly. After some length of time, the researchers will collect the data from the four groups and compare the results. With the only difference being which treatment was taken, researchers can make conclusions determining which treatment (if any) worked better than the others.



There are some other potential problems here though. For instance, would you want the subjects to know which medication they are receiving? It is very possible that they may have some preconceived notions regarding the effectiveness of one or more of these medicines. Such unconscious bias can influence how they perceive the treatment to work. What researchers often do to avoid any bias that the subjects will bring with them is to not tell them what treatment they are receiving. Such an experiment is said to be **blind**. It is also possible that the researcher may have preconceived notions, or hopeful expectations, regarding the effectiveness of one or all of the treatments. To avoid this, a third party can package the various treatments in similar looking containers, each marked only with a code, before the researcher distributes them to the subjects. In this case neither the subjects nor the researcher distributing the treatments know who is getting what. This is a **double blind** experiment, and is used often in clinical trials to limit bias.

Another issue is that often a patient's symptoms may improve just at the 'idea' of getting a medication. This is called the **placebo effect**. Imagine a child who is crying dramatically over a scraped knee, but stops immediately once mom puts a bandaid on. The bandaid is the placebo. It is also common for a participant, who believes that she or he is receiving a potentially promising medication, to have symptoms improve simply because of her or his expectation that they will. To account for this placebo effect, researchers will often give the control group a fake treatment called a **placebo**. A placebo is sometimes called a "sugar pill"– it looks like the real treatment, but has no active ingredients. Placebos make blind and double-blind experiments possible. An experiment could involve a placebo shot, or even a placebo surgery (aka sham surgery).

We will demonstrate how to **outline an experiment** through the following examples. See the sample outline above as a reference.

Example 1

Suppose that a group of scientists have developed a medication that they believe will cure mean-ness. They are calling it Kind At Last (KAL). There are 520 mean people who are willing to participate in this study (300 males and 220 females). This pill needs to be taken twice daily and it may take a few weeks to be fully absorbed into a person's system. Identify the following, and outline a completely randomized experiment.

a) Subjects

b) Explanatory Variable

c) Response Variable

d) Will it be blind? Double-blind? placebo controlled? is a pre-test necessary?

e) Outline a completely randomized experiment

Solution

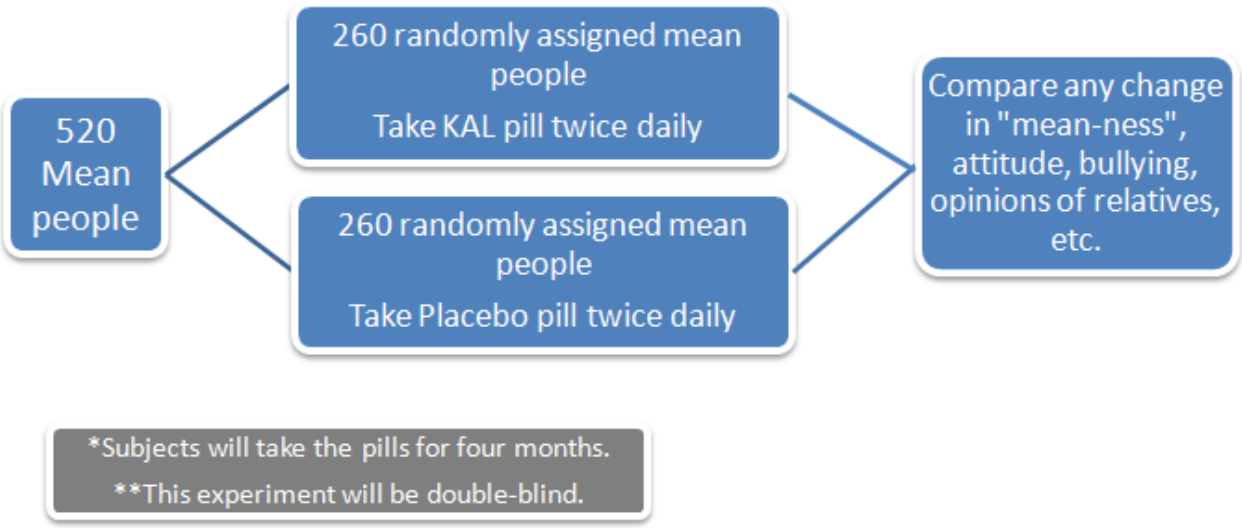
a) Subjects: the 520 mean people (330 male & 220 female)

b) Explanatory Variable: the KAL pills

c) Response Variable: any change in mean-ness

d) will it be blind? double-blind? placebo controlled? is a pre-test necessary? this could definitely be placebo controlled and double-blind. Neither the patients, nor the person distributing the medicine will know which people are receiving which medication. The KAL pills and the placebos will look identical and be in similar packages.

e) Outline a completely randomized experiment:


520 Mean people
260 randomly assigned mean people
Take KAL pill twice daily
260 randomly assigned mean people
Take Placebo pill twice daily
Compare any change in "mean-ness", attitude, bullying, opinions of relatives, etc.
*Subjects will take the pills for four months.
**This experiment will be double-blind.

The previous example is the a **completely randomized experiment** because all of the subjects started in one group. All subjects were then randomly assigned to treatment groups, with any combination of genders being possible. What if it was theorized that this medication actually has different effects on males than on females? With a completely randomized design it is very possible that we would not end up with an equal number of males and females in each treatment group. If that were to happen, we would not be able to tell whether the treatment affected different genders in the same way or not.

Randomized Block Designs

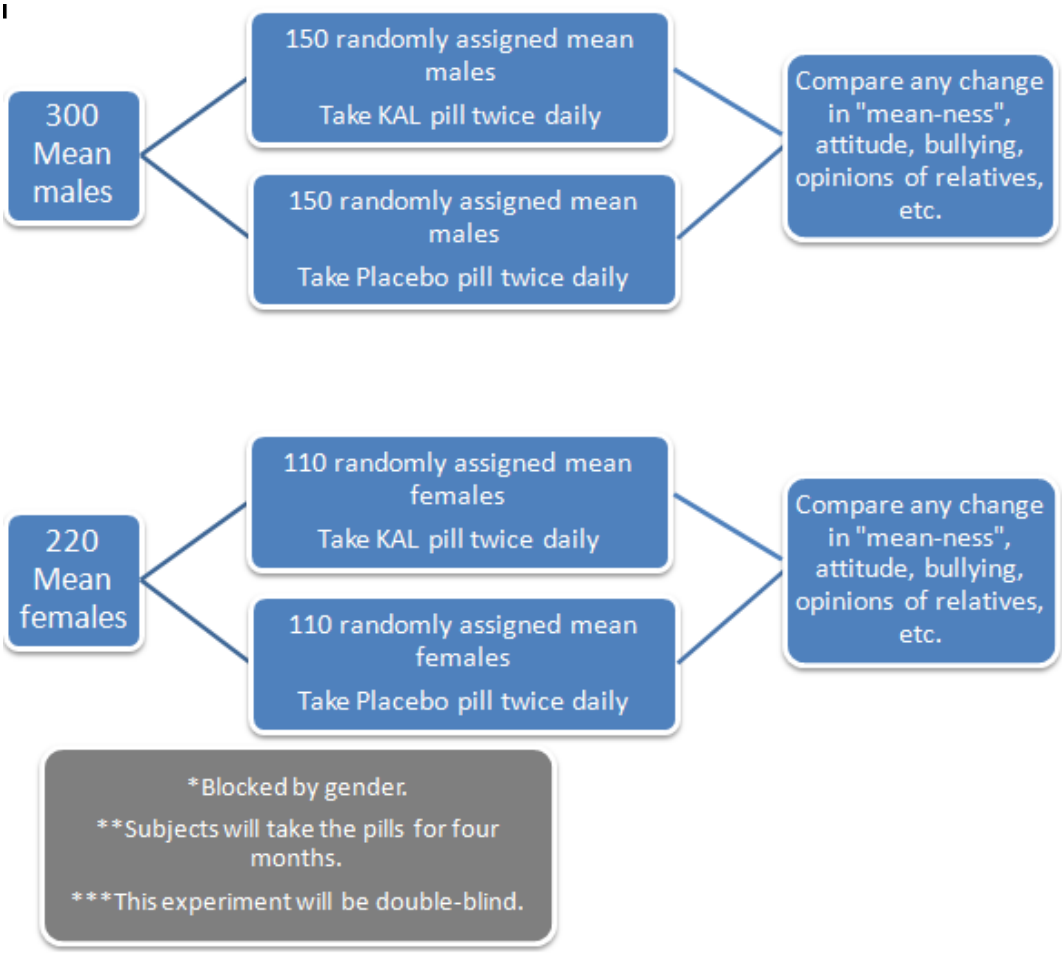
In such a case, it is a good idea to involve blocking in your experimental design. When it is suspected that different subgroups may respond differently to the treatment, the statisticians separate them at the beginning into intentional subgroups called **blocks**. The subjects in an experiment may be blocked by age, gender, race, previous medical history, etc. Be sure that you do not say that you will randomly assign to the blocks. You cannot randomly choose who is male or female, and you cannot randomly choose who is which race, etc. Each block is then randomly divided among the various treatment groups. This assures a more equal distribution of the subjects among the treatments. It also directly addresses the effects of this suspected lurking variable. Experimental designs in which blocking is used are called **randomized block designs**.

Example 2

Outline a randomized block design to test the KAL pills that blocks by gender. (Continued from example 1)

Solution

Outline a randomized block design–


300 Mean males
150 randomly assigned mean males
Take KAL pill twice daily
150 randomly assigned mean males
Take Placebo pill twice daily
Compare any change in "mean-ness", attitude, bullying, opinions of relatives, etc.
220 Mean females
110 randomly assigned mean females
Take KAL pill twice daily
110 randomly assigned mean females
Take Placebo pill twice daily
Compare any change in "mean-ness", attitude, bullying, opinions of relatives, etc.
*Blocked by gender.
**Subjects will take the pills for four months.
***This experiment will be double-blind.

Once you have done the comparisons within blocks, you will also want to compare across blocks to see if there are differences. For example, perhaps this KAL medicine works really well on males, but doesn't do a thing for females.

Section 4.5 Exercises

1. Researchers want to determine how effective a new allergy drug called Scratch-Be-Gone is at reducing pet allergies. One pill should be taken daily with a meal. 450 pets suffering from allergies will participate in a clinical study comparing this new drug with an existing market drug and a placebo. Identify each of the following:

a) Subjects

b) Explanatory Variable

c) Response Variable

d) Is it possible for this experiment to be double-blind? Explain.

e) Outline a completely randomized experiment

2. Ms. Rokinroll has a theory that listening to music while working on probability problems will help students retain knowledge. She has a set of earphones for each student and intends to compare the effects of classical music, country music, and heavy metal music. Her first period probability class has 36 students and her last period probability class has 34 students. Identify each of the following:

a) Subjects

b) Explanatory Variable

c) Response Variable

d) Do you feel that a control group of no music is necessary? Why or why not?

e) Do you feel that any of the following should be a part of this experiment: blind, double-blind, pre-test? placebo controlled?

f) Outline a randomized block design experiment

3. Researchers want to test a new eye drop against Blink Brand Eye Drops to see if it is better at reducing dry eye symptoms for contact wearers. The researchers are also interested in whether males and females will respond differently. The subjects available are 480 male and 502 female contact wearers who suffer from frequent dry eyes. Identify the following:

a) Subjects

b) Explanatory Variable

c) Response Variable

d) Outline an appropriate experiment: (*will it be blind? double-blind? blocked? placebo controlled?*)

e) Clearly explain how a table of random digits can be used to do the randomization. Using line #129 select the first five males who will be in the first treatment group.

4. A new type of cell phone is being developed by The Millionaire Phone Makers Corporation. This phone, called *Make-Us-More-Money (MUMM),* has a target audience of college students and young professionals (18-35 year olds). The company has developed three different ad campaigns – a commercial for each has been made and will be tried on the test subjects. The company wants to determine which ad campaign will be most effective prior to flooding the market, so they will test the various commercials on 744 University of Wisconsin students, and 3,057 people who attend this year's Young Professionals Conference in Los Angeles. After viewing a commercial, each subject will fill out a questionnaire that test how likely they would be to purchase the *MUMM* phone. Identify each of the following:

a) Subjects

b) Explanatory Variable

c) Response Variable

d) Outline an appropriate experiment (it will need to be blocked by the two different locations)

e) This scenario is different than the previous examples, because in the other examples we were able to do the randomization in advance. That would not be possible for something like this in both locations, the subjects will walk up to the researcher and need to be assigned to a 'treatment group'. Explain how the randomization can be done in this case.

4.6 Chapter 4 Review

This chapter covers the topics of data collection methods and potential sources of bias. We learned about experiments, observational studies, sample surveys and censuses. Several potential errors and sources or bias were introduced. We also learned how to use a random digit table to make random selections, how to outline experimental designs, and how to calculate and state 95% confidence intervals.

You should go back and read through each of the sections in this chapter, paying careful attention to all of the new terms in bold. This will help you to do problem 1 from your homework assignment.

Review Exercises

1) Study your new vocabulary!

a) Make flashcards with all of the terms from this chapter. Write the term on one side of the card. On the other side, write a brief definition and include an example.

b) Study your flashcards.

2) Each statement below claims that the ACT's are not a fair measurement for college readiness, but for a different reason. For each student's statement, determine whether he or she is questioning the validity, the reliability, or claiming that it will be biased. Explain your answers.

a) *The ACT's are not fair because it is timed and I cannot work fast enough. Consequently I am not really doing as well as I could; I always get a lower score than I should receive.*

b) *The ACT's are not fair because the vocabulary is not clear and I do not even understand what the questions are asking me. I always study really hard and do my homework and I am totally ready for college, but that doesn't show up on some stupid test.*

c) *The ACT's are not fair because the first time I took them I scored a 21, but the second time I scored a 16. How can that be right?*

3) Suppose you want to take a simple random sample of 350 women, from a population of 4700 females on the University of Coolness campus.

a) Explain the steps you would follow if you were going to make the selection using a table of random digits (be thorough).

b) Starting at line 137, select the first five numbers.

4) Suppose that after carrying out your survey of 350 women, you found that 74 of the women said that they "did not feel safe walking on campus after dark". Identify each of the following.

a) Population of interest

b) Parameter of interest

c) Sample

d) Statistic

e) Margin of error (quick method for 95%)

f) Calculate a 95% confidence interval

g) Make a 95% confidence statement

5) A high school social studies teacher wants to see if giving a completely multiple choice test versus a traditional free response test will improve student scores. She has designed two versions of the chapter 6 test to test her question. She has two classes – a 1st hour class with 32 students, and a 5th hour class with 37 students. The two classes are very different in both behavior and academic performance, so she decides to carry out her experiment using blocking. Identify the following:

a) Subjects

b) Explanatory variable(s)

c) Response variable

d) Will this experiment be blind? double-blind? placebo controlled?

e) Outline a randomized block experiment.

6) Suppose that you are trying to determine whether kindergarten students who have gone to child care centers show more aggressive behaviors than children who have not attended child care centers. You have the data regarding whether or not each child attended a child care center and for what length of time. You are now going to study aggressive behaviors. For each of the following, decide which type of data collection method is being proposed: observational study, sample survey, census, or experiment.

a) Observers will watch the kindergartners on the playground, recording aggressive behaviors.

b) A survey will be given to 20 randomly selected parents asking each to rate his or her child's behavior.

c) A survey will be given to all kindergarten parents asking each to rate his or her child's behavior.

d) During center time, a teacher will take a toy away from a child and record whether they act aggressively.

7) For the study in question 6

a) Suggest something that may go wrong with, or may be a source of bias, for each of the proposed data collection methods.

b) Which of the methods do you feel will yield the best results? Explain.

8) A researcher at the University of Minnesota believes that a certain component of ant venom can be used to lessen the amount of swelling in the knuckles of people suffering from arthritis. The ant venom treatment has been made into a capsule form that can be swallowed, and is designed to be taken one time per day. Suppose that you have 200 people suffering from arthritis who have volunteered to participate in this study. Identify the following:

a) Subjects

b) Explanatory variable

c) Response variable

d) Will this experiment be blind? double-blind? placebo controlled?

e) Outline a completely randomized design

9) Your teacher wants to find out whether chocolate helps students concentrate on their tests. In one class, she gives all of the students chocolate before the test begins. In another class, she does nothing. Is this an example of an observational study, a sample survey, a census, or an experiment? Give reasons to support your answer.

10) You bought a sweater on discount that was originally marked at $30. When you got to the register, it rung up as $23. What was the percent discount?

11) The cost of gas in 2001 was $1.45 per gallon. The average cost today is $3.75.

a) What is the amount of increase?

b) What is the percent of increase?

12) The table below shows the number of seniors and the number of seniors graduating for three high schools.

School	Number of Seniors	Number Graduating	Graduation Rate %
McArthur	423	354	
Meade	125	110	
Eisenhower	392	379	

a) Which school has the most students graduating?

b) Determine the graduation rate for each school.

c) Which school has the highest graduation rate?

13) Identify the sampling method used in each of the following: SRS, stratified random sample, systematic random sample, multi-stage random sample, random cluster sample, voluntary response sample, convenience sample

a) Every fifth person boarding a plane is searched thoroughly.

b) At a local community College, five math classes are randomly selected out of 20 and all of the students from each class are interviewed.

c) A researcher randomly selects and interviews forty male and forty female teachers, from a university with 122 female and 135 male instructors.

d) A researcher for an airline interviews all of the passengers on five randomly selected flights.

e) Based on 12,500 responses from 42,000 surveys sent to its alumni, a major university estimated that the annual salary of its alumni was 92,500.

f) A community college student interviews everyone in his biology class to determine the percentage of students that own a car.

g) A market researcher randomly selects 200 drivers under 35 years of age and 100 drivers over 35 years of age, from those insured with Quality Car Insurance.

h) A researcher selects 12 states randomly. From each state, she randomly selects 20 middle schools. From each middle schools, she randomly selects 15 teachers. The 3,600 teachers were then interviewed by phone..

i) To avoid working late, the quality control manager inspects the last 10 items produced that day.

j) The names of 70 contestants are written on 70 cards. The cards are placed in a bag, and three names are picked from the bag.

14) The athletic director wants to know how tax payers in the community feel about funding for athletics at the high school. He surveys his coaches and the parents of athletes at his school. Describe what is wrong with his methodology.

15) Suppose that a poll was commissioned to determine whether people in the U.S. believe that pro wrestling is a sport. Identify the type(s) of bias that will likely be present in each of the following scenarios. Some will have more than one. Explain your answers.

a) An online poll was sent to all visitors of the WWE website.

b) Telephone interviews are done to randomly selected phone numbers between 4 p.m. and 8 p.m.

c) Some people were embarrassed to admit that they liked wrestling and think it is a sport.

d) One of the questions asked was, "Do you believe that the pro wrestlers are actors or should they be considered serious athletes?"

e) One of the researchers spilled coffee on a big stack of surveys and several had to be thrown away.

f) A second poll done had slightly different results.

g) WWE had fans fill out a survey as they left a pro wrestling event.

Picture of cell phones. http://www.w-cellphones.com July 19, 2011.

Students measuring heights. http://t3.gstatic.com July 25, 2011.

Pop can.http://popartmachine.com July 25, 2011.

Calvin and Hobbes. http://www.stat.psu.edu/old_resources/Cartoons/cartoon014.gif July 27, 2011.

Clipboard. http://boylston.bbrsd.schoolfusion.us July 25, 2011.

Hockey Players.http://images.paraorkut.com July 25, 2011.

Scientist:http://www.reversingibs.com July 15, 2011.

Chapter 5

Analyzing Univariate Data

Introduction

So now that we have discussed some methods for collecting data we can look at what to do with those findings. Whether you have collected categorical or numerical data you will want to choose an appropriate type of graphical display so that you can see the data. Charts and graphs of various types, when created carefully, can provide important information about a data set. You will also need to analyze the data with numerical and summary statistics. Once you have constructed a graphical display and have calculated numerical statistics, it will be necessary to describe your findings verbally. Statisticians, such as yourself, then make appropriate conclusions and comparisons based on the data and statistics, avoiding opinions and judgment statements. This chapter will concentrate on some of the more common visual presentations of data, numerical analysis of data, and verbal descriptions of data.

5.1 Categorical Data

- Organize categorical data in tables
- Construct bar graphs and pie charts by hand and with computer software programs
- Describe, summarize and compare categorical data

Each student in the class should complete the following survey. The data collected will be used in your homework problems. Notice that the variables in each question are categorical.

1. *What is your gender?* **Choose one**
   - *Female*
   - *Male*
2. *What is your favorite season?* **Choose one**
   - *Winter*
   - *Spring*
   - *Summer*
   - *Fall*
3. *Which of these is your favorite type of food?* **Choose one**
   - *Italian*
   - *Asian*
   - *Mexican*
   - *American*
4. *What type of pet(s) do you have?* **Choose all that apply**
   - *Dog*
   - *Cat*
   - *Fish*
   - *Reptile*
   - *Rodent*
   - *Other*
   - *None*

Frequency Tables and Bar Graphs




(Note- Two videos for this section)

When analyzing categorical data (also called qualitative data), bar graphs are commonly used. A **bar graph** is a graph in which each bar shows how frequently a given category occurs. It is usually helpful to organize the data in a **frequency table**, a table that shows the number of occurrences for each category, before constructing the bar graph. The bars can go either horizontally or vertically, they should be of consistent width, and need to be equally spaced apart. The categories are separate and can be put in any order along the axis. It is common to put them in alphabetical order, but not needed. And, as with all of the graphs you will construct, be sure to use a consistent scale, include a title, labels for axes, numbers to mark axes as necessary, and a key whenever needed.

Example 1

A bar graph could show the types of pets of a group of students for example. Here are the types of pets owned by a class of 33 geometry students.

a) Why do the numbers add up to more than 33?

b) Construct a bar graph to show this class' data.

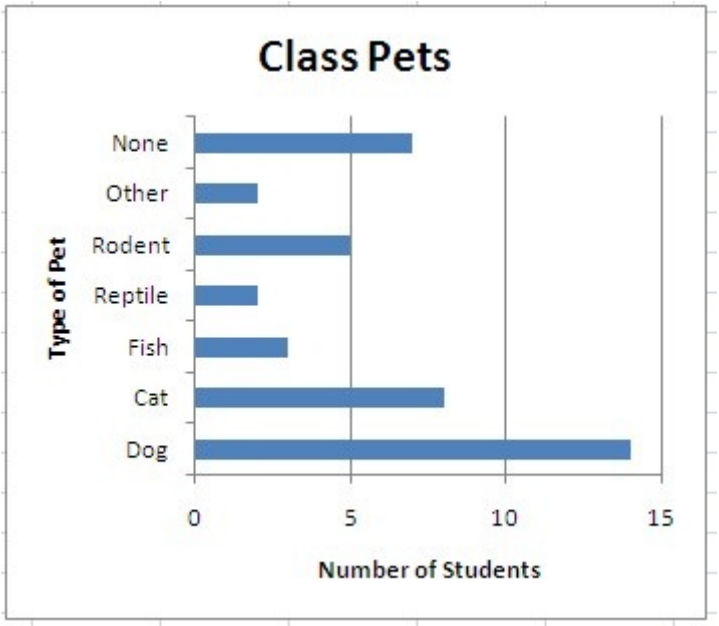
c) Describe what the graph shows.

Type of Pet	#of Students
Dog	14
Cat	8
Fish	3
Reptile	2
Rodent	5
Other	2
None	7

Solution

a) ***They add up to more than 33 because some students own more than one type of pet and are being counted in more than one category.***

b) Here is a bar graph that was created using Excel:



c) ***For this class, the most common pet is a dog. Fourteen students, or 42% of the class, own a dog. Having a cat, or no pet at all are the next most common events. Five students own some type of rodent, two have reptiles for pets, and three have fish. There are also two students who own some other type of pet.***

Example 2

A great deal of electronic equipment ends up in landfills as people update their computers, TVs, cell phones, etc. This is a concern because the chemicals from batteries and other electronics add toxins to the environment. This Electronic Waste has been studied in an effort to decrease the amount of pollution and hazardous waste. The following frequency table shows the amount of tonnage of the most common types of electronic equipment discarded in the United States in 2005. Construct a bar graph and comment on what it shows.

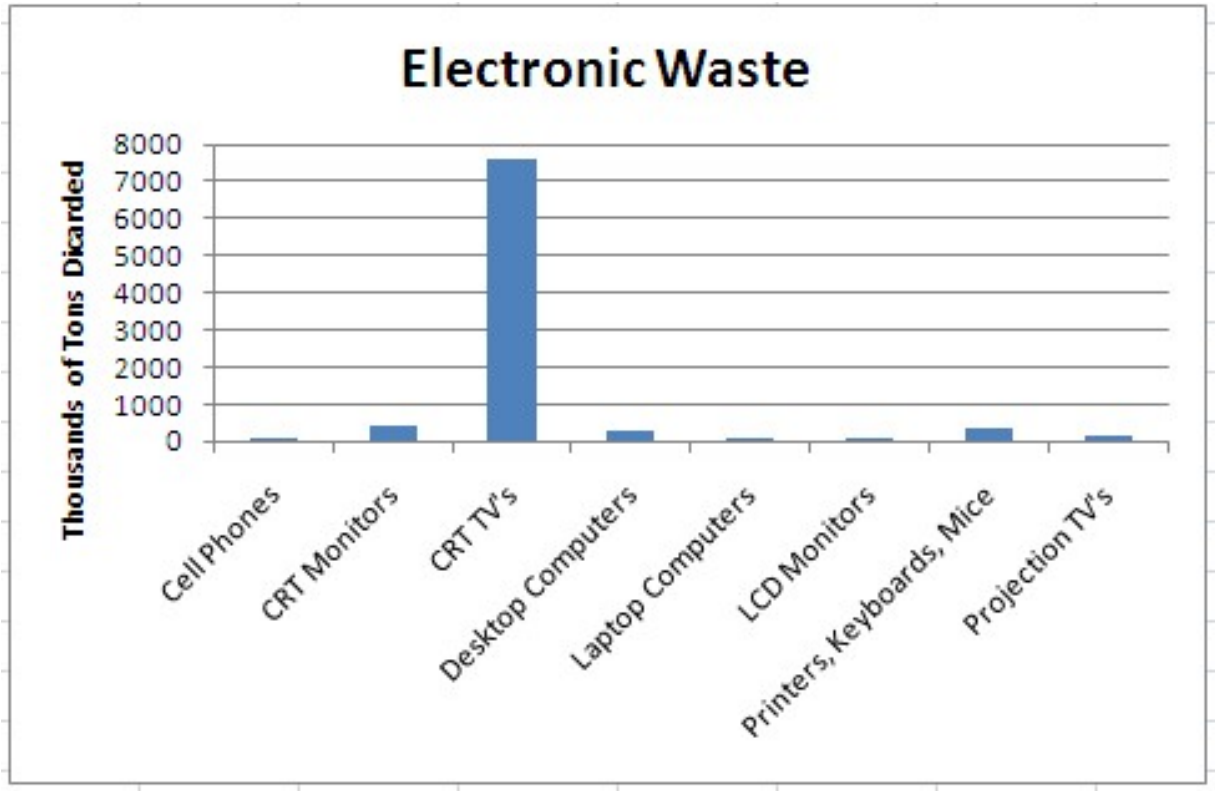
Table 5.1:

Electronic Equipment	**Thousands of Tons Discarded**
Cathode Ray Tube (CRT) TV's	7591.1
CRT Monitors	389.8
Printers, Keyboards, Mice	324.9
Desktop Computers	259.5
Laptop Computers	30.8
Projection TV's	132.8
Cell Phones	11.7
LCD Monitors	4.9

Electronics Discarded in the US (2005). *Source:* National Geographic, January 2008. Volume 213 No.1, pg 73.

Solution

The type of electronic equipment is a categorical variable, and therefore, this data can easily be represented using the *bar graph* below:



According to this 2005 data, the most commonly disposed of electronic equipment was CRT TV's, by more than 19 times that of the next type of electronic equipment.

Pie Charts



Pie charts (or circle graphs) are used extensively in statistics. These graphs are used to display categorical data and appear often in newspapers and magazines. A **pie chart** shows each category (sectors) as a part of the whole (circle). The relationships between the parts, and to the whole, are visible in a pie chart, by comparing the sizes of the sectors (slices). Constructing a pie chart uses the fact that the whole of anything is equal to 100%-all of the sectors equal the whole circle. Remember from geometry that the central angles of a circle total 360^0. So, in regard to pie charts, $360^0 \equiv 100\%$ of the circle . The sections should have different colors or patters to enable an observer to clearly see the difference in size of each section.

Pie charts are the appropriate choice when you are working with categorical data that covers 100%. It is not an appropriate choice when you aren't working with 100% or when choices may include overlaps. For

example, when we asked every student in this class to list the pets they currently have, we found some students who have more than one pet. So a pie chart would not be an appropriate way to display that data. The sectors in a circle graph do not allow for overlaps such as this. Another time when pie charts are not appropriate is when the choices do not cover all possibilities. For example, the electronic waste example above does not include every possibility, so the categories would not add to 100%. In such cases a bar graph would be a more appropriate choice, because it allows for overlaps and does not need to cover exactly 100% of the choices.

Example 3: How to Construct a Pie Chart

The Red Cross Blood Donor Clinic had a very successful morning collecting blood donations. Within three hours twenty-five people had made donations. The types of blood dontated are:

Table 5.2:

Blood Type	**A**	**B**	**O**	**AB**
Number of donors	7	5	9	4

Construct a pie chart to represent the data.

Solution

Step 1: Determine the total number of donors. $7 + 5 + 9 + 4 = 25$

Step 2: Express each donor number as a percent of the whole by using the formula $\boxed{Percent = \frac{f}{n} \cdot 100\%}$ where f is the frequency and n is the total number.

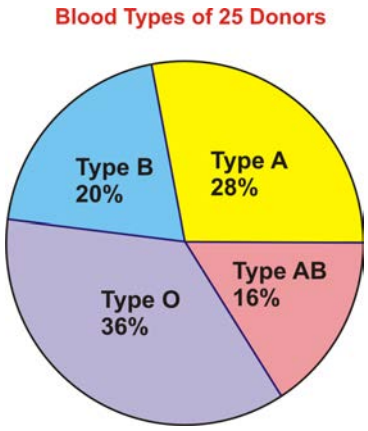
$$\frac{7}{25} \cdot 100\% = 28\% \quad \frac{5}{25} \cdot 100\% = 20\% \quad \frac{9}{25} \cdot 100\% = 36\% \quad \frac{4}{25} \cdot 100\% = 16\%$$

Step 3: Express each donor number as the number of degrees of a circle that it represents by using the formula $\boxed{Degree = \frac{f}{n} \cdot 360^\circ}$ where f is the frequency and n is the total number.

$$\frac{7}{25} \cdot 360^\circ = 100.8^\circ \quad \frac{5}{25} \cdot 360^\circ = 72^\circ \quad \frac{9}{25} \cdot 360^\circ = 129.6^\circ \quad \frac{4}{25} \cdot 360^\circ = 57.6^\circ$$

Step 4: Using a protractor or technology to make the central angles, graph each section of the circle.

Step 5: Write the label and correct percentage inside the section. Color each section a different color. Be sure to include a title, and a key if needed.



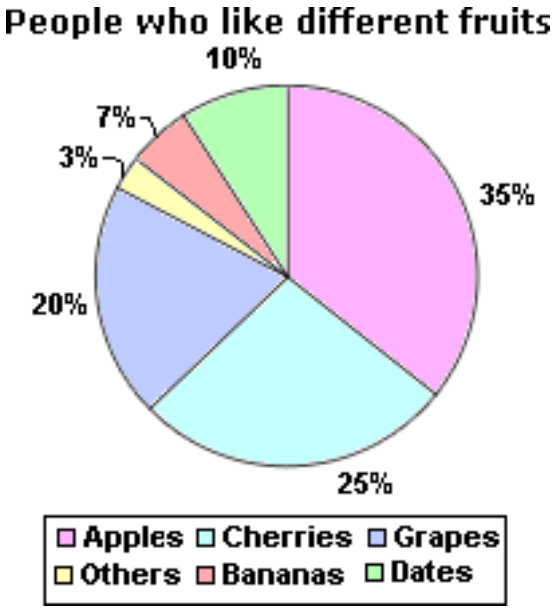
From the graph, you can see that more donations were of Type O than any other type. The least amount of blood collected was of Type AB. In order to create a pie graph by using the circle, it is necessary to use the percent of a section to compute the correct degree measure for the central angle. The blood type graph labels each section with context and percent, and not the degrees. This is because degrees would not be meaningful to an observer trying to interpret the graph. If the sections are not labeled directly as they are in that example, it is necessary to include a key so that the observers will know what each section represents.

Graphs on Computer Software

The above pie chart could be created by using a protractor and graphing each section of the circle according to the number of degrees needed for each section. However, bar graphs and pie charts are most frequently made with computer software programs such as Excel or Google Docs, if you would like to learn how to do this on Excel, click here. You will be asked to create bar graphs and pie charts using computer software. When you do this, be sure to include titles, labels, and keys when needed. Be sure to 'fix' the graph generated by the software program so that it looks the way you want it to look and shows clearly what ever it is you are trying to convey.

Example 4

Comment on what the graph shows:



Solution

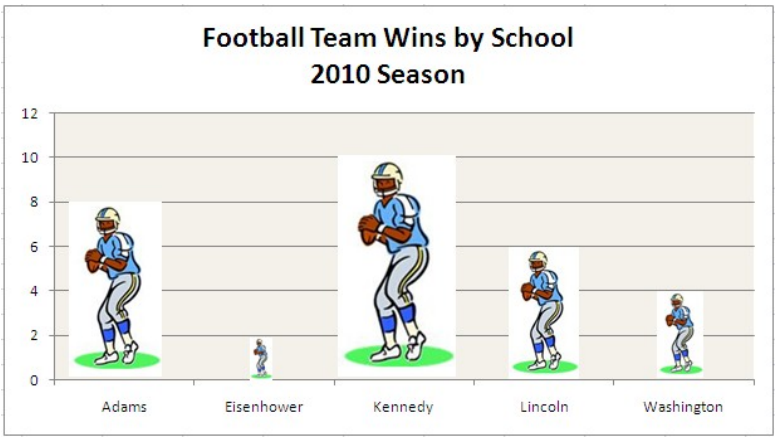
Several people were asked to choose their favorite fruits from a list of six options. Apples were the favorite choice with 35% of the participants choosing them. The second favorite fruit was cherries at 25%, followed by grapes with 20%. Ten percent of the people said that dates were their favorite fruit. However, only 7% chose bananas from the choices provided and the remaining 3% liked some fruit other than those listed.

Pictographs

Another type of graph that is sometimes used to display categorical data is a pictograph. A **pictograph** is basically a bar graph with pictures instead of bars. A problem with pictures in graphs is that the area that they take up can mislead the observer. The width and height both increase as the picture gets larger. Pictographs are often used, but should generally be avoided in serious statistical representations.

Example 5

The following graph compares the number of wins for high school football teams during the 2010 seasons. Explain why the pictograph is misleading.



Solution

The pictures increased in both height and width. So when something should be doubled, it actually looks four times as big. For example, when comparing the number of wins between Eisenhower and Adams the graph should show 4 times as many wins. However, in this pictograph it looks as though Adams had 16 times as many wins (4 times as wide X 4 times as tall).

Section 5.1 Exercises

1) Computer equipment contains many elements and chemicals that are either hazardous, or potentially valuable when recycled. The following data set shows the contents of a typical desktop computer weighing approximately 27 kg. Some of the more hazardous substances, like Mercury, have been included in the "other" category, because they occur in relatively small amounts that are still dangerous and toxic.

Table 5.3:

Material	**Kilograms**
Plastics	6.21
Lead	1.71
Aluminum	3.83
Iron	5.54
Copper	2.12
Tin	0.27
Zinc	0.60
Nickel	0.23
Barium	0.05
Other elements and chemicals	6.44

Figure: Weight of materials that make up the total weight of a typical desktop computer. *Source:* http://dste.puducherry.gov.

a) Create a bar graph for this data.

b) Complete the chart below to show the approximate percentage of the total weight for each material and the central angle needed.

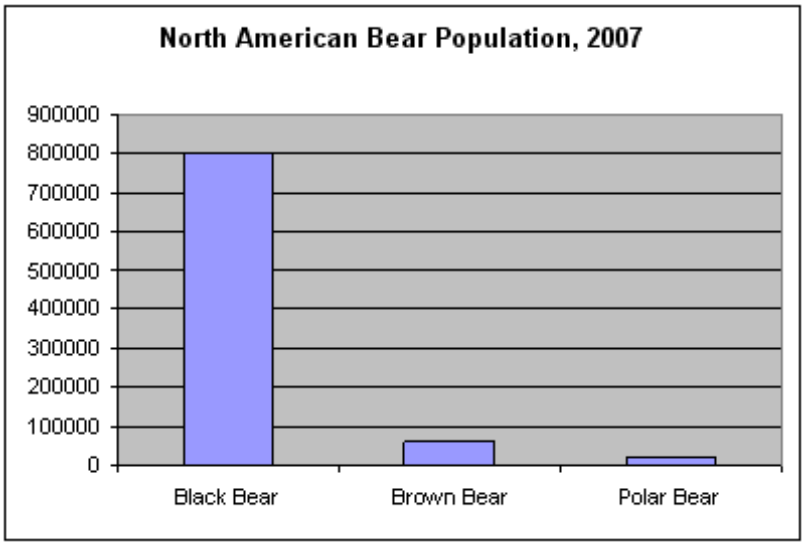
c) Why is a pie chart also appropriate for this example?

d) Create a pie chart for this data.

Table 5.4:

Material	**Kilograms**	**Approximate % of Total Weight**	**Central Angle for Pie Graph**
Plastics	6.21		
Lead	1.71		
Aluminum	3.83		
Iron	5.54		
Copper	2.12		
Tin	0.27		
Zinc	0.60		
Nickel	0.23		
Barium	0.05		
Other elements	6.44		

2) Based on what you can see in the graph, write a brief description of what it is showing. This should be at least three sentences and in context.



Source.http://www.mathworksheetscenter.com Aug. 5, 2011.

3) Type of Pet?

a) Construct a frequency table to show the Type of Pet data from our class.

b) Use Excel or Google Docs to create a bar graph that shows the types of pets the students in our class have.

c) Write a brief description of what your graph shows.

4) Favorite Season?

a) Construct a frequency table to show the Favorite Season data from our class.

b) Use Excel or Google Docs to create a pie chart that shows the favorite season of the year for the students in our class.

c) Write a brief description of what your graph shows.

5) Look at the school lunch graph that was created by some students:



a) In what way is this graphical representation misleading? Explain.

b) Create a better graphical representation for this same data.

6) Favorite Foods?

a) Construct a frequency table to show the Favorite Food data **separately** for males and females for our class.

b) Use Excel or Google Docs to create two pie charts that compare the favorite food types for the boys and girls in our class. The charts should 'match' as much as possible– they should be the same size and use the same colors, fonts, etc.

c) Write a brief description comparing the boys and girls choices for favorite food. Look for similarities and differences.

7) The following table has Minnesota Wild statistics for 2010-2011, for some of the Wild players. Thirteen variables are listed across the top and have been highlighted.

a) Identify the individuals.

b) Identify what each variable is (Example GP = games played). You may need to do some research.

c) Classify each variable as numerical or categorical?

2010-2011 REGULAR SEASON

Forwards & Defensemen

#	POS	PLAYER	GP	G	A	P	+/-	PIM	PP	SH	GW	S	S%
24	R	MARTIN HAVLAT	78	22	40	62	-10	52	3	0	4	229	9.6
9	C	MIKKO KOIVU	71	17	45	62	4	50	7	1	3	191	8.9
15	L	ANDREW BRUNETTE	82	18	28	46	-7	16	8	0	3	117	15.4
8	D	BRENT BURNS	80	17	29	46	-10	98	8	0	3	170	10.0
7	C	MATT CULLEN	78	12	27	39	-14	34	5	4	2	150	8.0
96	C	PIERRE-MARC BOUCHARD	59	12	26	38	-3	14	0	0	2	98	12.2
21	C	KYLE BRODZIAK	80	16	21	37	-4	56	2	1	1	126	12.7
20	R	ANTTI MIETTINEN	73	16	19	35	-3	38	8	0	4	168	9.5
22	R	CAL CLUTTERBUCK	76	19	15	34	-5	79	4	0	3	191	9.9

5.2 Time Plots & Measures of Central Tendency



Learning Objectives

- Construct time plots
- Describe trends in time plots
- Calculate range and measures of central tendency: mean, median, mode
- Understand how a change in the data will effect the statistics

Line Graphs as Time Plots

We are often interested in how something has changed over time. The type of graphical display that shows this the most clearly is a **time plot**, or line graph. When one of the variables is time, it will almost always be plotted along the horizontal axis (as the explanatory variable). Because time is a continuous variable and we are trying to see if there is some type of trend in how the other variable (response) has behaved over a period of time, a line graph is often very useful in showing this relationship.

Example 1

The total municipal waste generated in the US by year is shown in the following data set.

a) Construct a time plot to show the change in the amount of municipal waste generated in the United States during the 1990's.

b) Comment on the trend that is shown in the graph.

c) Suggest factors (other than time) that may be leading to this trend.

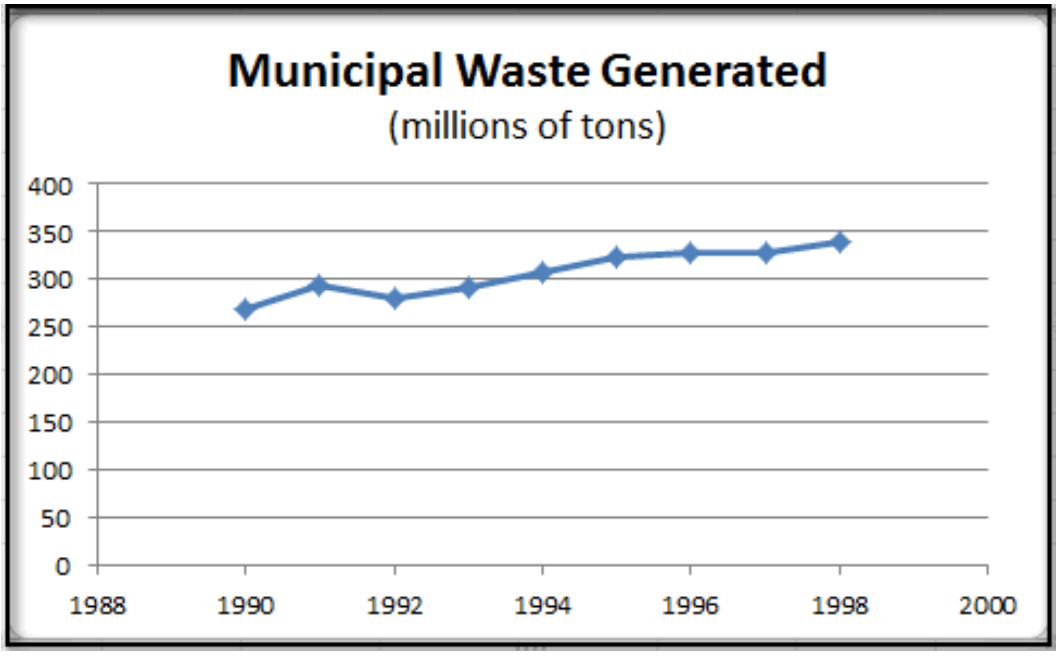
Table 5.5:

Year	**Municipal Waste Generated** (Millions of Tons)
1990	269
1991	294
1992	281
1993	292
1994	307
1995	323
1996	327
1997	327
1998	340

Source: http://www.zerowasteamerica.org

Solution

a) In this example, the time (in years) is considered the *explanatory variable*, and is graphed along the horizontal axis. The amount of municipal waste is the *response variable*, and is graphed along the vertical axis. Time plots can be drawn by hand, graph paper makes this easier, or created with computer software programs, or graphing calculators. This example was made using Excel.



b) This graph shows that the amount of municipal waste generated in the United States increased at a fairly steady rate during the 1990s. Between 1991 and 1992 there was a decrease of 13 million tons of municipal waste, but every other year during the 1990s had an increase.

c) It should be noted that factors other than the passage of time cause our waste to increase. Other factors, such as population growth, economic conditions, and societal habits and attitudes also contribute as causes.

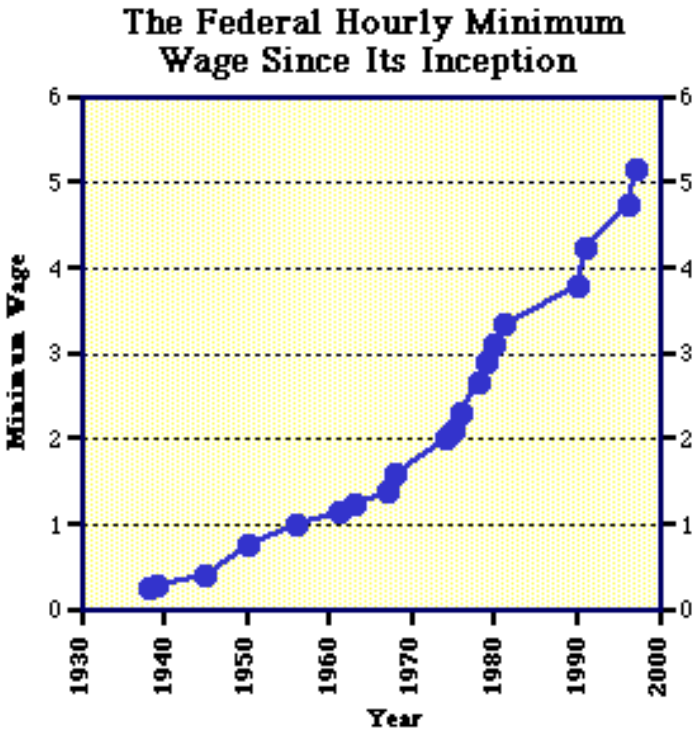
Example 2

Here is a line graph that shows how the hourly minimum wage changed from when it was first mandated through 1999.

a) During which decade did the hourly wage increase by the greatest amount?

b) During which decade did it increase the most times?

c) When did it stay constant for the longest?



Source.**http://mste.illinois.edu** Aug 1,2011.

Solution

a) The greatest increase appears to have happened during the 1990s, when it went from ≈$3.75 to ≈$5.20.

b) The 1970s appear to have had 5 or 6 increases in the minimum wage.

c) The longest constant minimum wage was during the 1980s.

Measures of Central Tendency & Spread



Mean

The **mean**, often called the 'average' of a numerical set of data, is the sum of all of the numbers divided by the number of values in the data set. This value is the arithmetic mean, and it tells us what value we would have if all of the data were the same. The mean is the *balance point* of a distribution, and is one of the three measure of central tendency commonly used in statistics. The mean, the median, and the mode

are all measures of central tendency. They all show where the center of a set of data "tends" to be. Each one is useful at different times. The mean is a summary statistic that gives you a description of the entire data set and is especially useful with large data sets where you might not have the time to examine every single value. However, the mean is affected by extreme values, called outliers, and can end up leaving the observer with the wrong impression of a data set. Imagine if we were to report the average salary of all employees of Burger Boy, including the main manager, the mean salary would give the impression that pay at Burger Boy is pretty high. But, if we removed the manager's salary from the calculation we would see a much different 'average' salary.

Median

The **median** is the number in the middle position once the data has been organized. Organized data is simply the numbers arranged from smallest to largest or from largest to smallest. This is the only number for which there are as many above it as below it in the set of organized data, and is referred to as the *equal areas point*. The median, for an odd number of data, is the value that is exactly in the middle of the ordered list, it divides the data into two halves. The median for an even number of data, is the mean of the two values in the middle of the ordered list. The median is useful when there are a few extreme values that can effect the mean, because the middle number will stay in the middle. The median often gives a good impression of the center, because there are 50% of the values above the median, 50% of the values below the median, and it doesn't matter how big the biggest values are or how small the smallest values are.

Mode

The **mode** of a set of data is simply the number that appears most frequently in the set. There are no calculations required to find the mode of a data set. You simply need to look for it. However, be aware that it is common for a set of data to have no mode, one mode, two modes or more than two modes. If there is more than one mode, simply list them all. And, if there is no mode, write 'no mode'. No matter how many modes, the same set of data will have only one mean and only one median. The mode is a measure of central tendency that is simple to locate but is not used much in practical applications. It is the only one of these three values that can be for either categorical or numerical data. Remember the example regarding pets? The mode was 'dogs' because that was the most common response.

Range

The range of a data set describes how spread out the data is. To calculate the **range**, subtract the smallest value from the largest value (maximum value – minimum value = range). This value provides information about a data set that we cannot see from only the mean, median, or mode. For example, two students may both have a quiz average of 75%, but one of them may have scores ranging from 70% to 82% while the other may have scores ranging from 24% to 90%. In a case such as this, the mean would make the students appear to be achieving at the same level, when in reality one of them is much more consistent than the other.



Example 3

Stephen has been working at Wendy's for 15 months. The following numbers are the number of hours that Stephen worked at Wendy's during the past seven months:

$$24, 24, 31, 50, 53, 66, 78$$

What is the mean number of hours that Stephen worked per month?

Solution

Stephen has worked at Wendy's for 15 months but the numbers given above are for seven months. Therefore, this set of data represents a sample of the population. The formula that is used to calculate the mean for a sample and for a population is the same. However, the symbols are different. The mean of a sample is denoted by $\bar{x}$ which is called "x bar". The mean of an entire population is denoted by μ which is the Greek letter "mu" (pronounced "myoo").

The number of data for a sample is written as n. The following formula represents the steps that are involved in calculating the mean of a sample:

$$Mean = \frac{add\ the\ numbers}{the\ number\ of\ numbers}$$

This formula can now be written using symbols.

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

You can now use the formula to calculate the mean of the hours that Stephen worked.

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$
$$\overline{x} = \frac{24 + 25 + 33 + 50 + 53 + 66 + 78}{7}$$
$$\overline{x} = \frac{329}{7}$$
$$\overline{x} = 47$$

The mean number of hours that Stephen worked during this time period was 47 hours per month.

Example 4

The ages of several randomly selected customers at a coffee shop were recorded. Calculate the mean, median, mode, and range for this data.

$$23, 21, 29, 24, 31, 21, 27, 23, 24, 32, 33, 19$$

Solution

mean: $(23 + 21 + 29 + 24 + 31 + 21 + 27 + 23 + 24 + 32 + 33 + 19)/12 = 307/12$

$307/12 = 25.58$

median: first, organize the ages in ascending order $19, 21, 21, 23, 23, 24, 24, 27, 29, 31, 32, 33$

second, count in to find the middle value $24, 24$ the middle value will be halfway between these two values (or the average of these two values) $\frac{24+24}{2} = 24$

mode: look for the value(s) that occur most frequently $21, 23, 24$ this data set has three modes

range: subtract the smallest value from the largest value (max - min = range) $33 - 19 = 14$

Solution: make your conclusion in context

At this coffee shop, the mean age of people in this sample is 25.58 years old and the median age is 24 years old. There were three modes for age at 21, 23, and 24 yeas old and the range for ages is 14 years.



Example 5

Lulu is obsessing over her grade in health class. She just simply cannot get anything lower than an A-,or she will cry! She knows that the grade will be based on her average (mean) test grade and that there will be a total of six tests. They have taken five so far, and she has received 85%, 95%, 77%, 89%, and 94% on those five tests. The third test did not go well, and she is getting worried. The cutoff score for an A is 93%, and 90% is the cuttoff score for an A-. She wants to know what she has to get on the last test. The teacher assures her that she will round to the nearest whole percent.

a) What is the lowest grade Lulu will need to get on the last test in order to get an A in health?

b) What is the lowest grade Lulu will need to get on the last test in order to get an A- in health?

Solution

a) So she sets up an equation thinking about how she would calculate her average test grade if she knew all six scores. Knowing that she wants the final average to equal 93%, she puts an '***x***' in the place of the last test score, and then does some algebra to solve for x.

$\frac{85+95+77+89+94+x}{6} = 93$

$(85 + 95 + 77 + 89 + 94 + x) = 93 * 6$

$85 + 95 + 77 + 89 + 94 + x = 558$

$440 + x = 558$

$x = 118$

Oh no! There is no way she can get 118%. So, there is no possible hope for her to get an A.

b) It is time to try for an A-, but that 118% scared her, so she is going to think of the lowest possible score that will still be an A-. With rounding, if she can get her mean score to 89.5%, she will make it. So she tries the same algebra, but with 89.5 as the final result.

$\frac{85+95+77+89+94+x}{6} = 89.5$

$(440 + x) = 89.5 * 6$

$440 + x = 537$

$x = 97$

There is hope! As long as she gets a 97% or higher on this last test, she can get an A-. She is going to study like crazy!



Section 5.2 Exercises

1) Determine the mean, median, mode and range for each of the following sets of numbers:

a) 20, 14, 54, 16, 38, 64

b) 22, 51, 64, 76, 29, 22, 48

c) 40, 61, 95, 79, 9, 50, 80, 63, 109, 42

2) The mean weight of five men is 167.2 pounds. The weights of four of the men are 158.4 pounds, 162.8 pounds, 165 pounds and 178.2 pounds. What is the weight of the fifth man?

3) The mean height of 12 boys is 5.1 feet. The mean height of 8 girls is 4.8 feet.

a) What is the total height of the boys?

b) What is the total height of the girls?

c) What is the mean height of the 20 boys and girls all together?

4) The following data represents the number of advertisements received by ten families during the past month. Make a statement describing the 'typical' number of advertisements received by each family during the month. Be sure to include statistics to support your statement.

43 37 35 30 41 23 33 31 16 21

5) Mica's chemistry teacher bases grades on the average of each student's test scores during the trimester. Mica has been kind of slacking this year, but hasn't been too concerned because he knows that he will at least get the credit (60% = passing). However, his parents just informed him that he will not be allowed to use the car if he has any grade below a C (73%). Here are Mica's chemistry test scores for the first eight chapters:

$$10, 70, 71, 82, 65, 76, 58, 75$$

a) Calculate the mean, median, mode, and range for Mica's chemistry tests. What grade will Mica receive in chemistry based on this?

b) His teacher has decided that each student may retake any one of his or her tests in an effort to improve his or her grade. Mica jumps at this opportunity, studies chapter one for hours and retakes the test. To his, and his mother's delight, his 10% turns into a 70%!! Woo-hoo! Calculate the mean, median, mode, and range for Mica after this change. Which of these values changed? Which did not? What grade will Mica receive now?

c) If Mica continues to study and earns an 60% on the chapter 9 test and a 76% on the chapter 10 test, what will his final average be?

d) If Mica continues to study and earns an 85% on the chapter 9 test and a 90% on the chapter 10 test, what will his final average be?

6) **Deals on Wheels:** The following table lists the retail price and the dealer's costs for 10 cars at a local car lot this past year:





Table 5.6:

Car Model	**Retail Price**	**Dealer's Cost**	**Amount of Mark-Up**	**Percent of Mark-Up**
Nissan Sentra	$24,500	$18,750		
Ford Fusion	$26,450	$21,300		
Hyundai Elantra	$22,660	$19,900		
Chevrolet Malibu	$25,200	$22,100		
Pontiac Sunfire	$16,725	$14,225		
Mazda 5	$27,600	$22,150		
Toyota Corolla	$14,280	$13,000		
Honda Accord	$28,500	$25,370		
Volkswagen Jetta	$29,700	$27,350		
Subaru Outback	$32,450	$28,775		

a) Calculate the **amount** each car was marked up.

b) Calculate the **percent** that each car was marked up $\frac{mark-up}{dealer-cost} * 100\%$ = Report answers rounded to the nearest one-tenth of a percent.

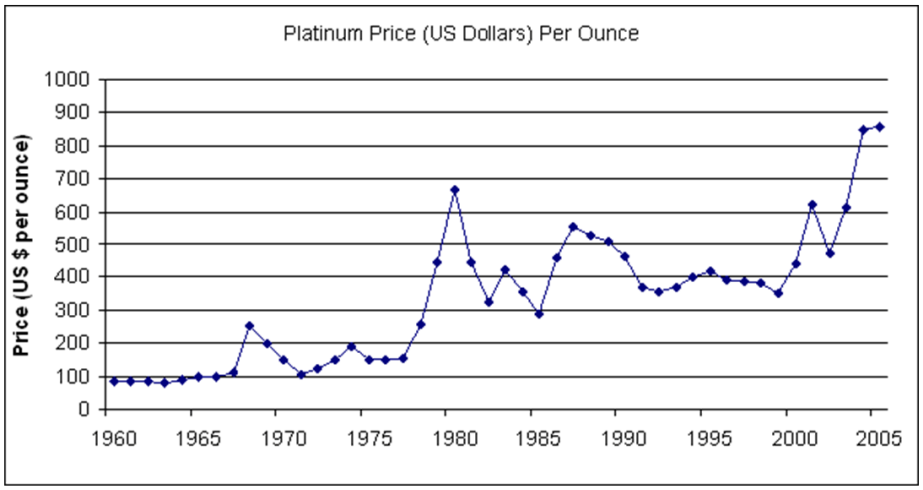
c) Calculate the mean, median, mode and range **for the percent of mark-up**.

d) Do the "amount of mark-up column" and the "percent of mark-up column" put the cars in the same order for profit? Explain or give an example.

7) Write a brief description of what the line graph for Platinum Prices shows. Be sure that you do this in context, as complete sentences, and that you include at least three observations.

Line Graph: Platinum Prices, 1960 to 2005

The line graph shows the price of platinum per ounce in US dollars between 1960 and 2005



Source.http://www.admc.hct.ac.ae

8) According to the U.S. Census Bureau, *"household median income"* is defined as *"the amount which divides the income distribution into two equal groups, half having income above that amount, and half having income below that amount."* The table shows the median household income data, every 3 years, from 1975 until 2008, according to the U.S. Census Bureau.

1975	1978	1981	1984	1987	1990	1993	1996	1999	2002	2005	2008
$11,800	$15,064	$19,074	$22,415	$26,061	$29,943	$31,241	$35,492	$40,696	$42,409	$46,326	$50,303

a) Construct a time plot for the median household data. You may do this by hand on graph paper, or by using technology.

b) Write a brief description of what the line plot shows. This should be done as complete sentences, in context of the distribution, and should include at least three distinct observations.



5.3 Numerical Data: Dot Plots & Stem Plots

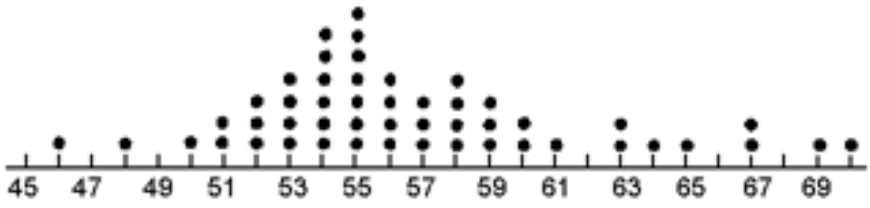
Learning Objectives

- Construct dot plots, stem plots and split-stem plots
- Calculate numerical statistics for quantitative data
- Identify potential outliers in a distribution
- Describe distributions in context– including shape, outliers, center, and spread

Dot Plots

One convenient way to organize numerical data is a dot plot. A **dot plot** is a simple display that places a dot (or X, or another symbol) above an axis for each datum value (datum is the singular of data). The axis should cover the entire range of the data, even numbers that will have no data marked above them should be included to show outliers or gaps. There is a dot for each value, so values that occur more than once will be shown by stacked dots. Dot plots are especially useful when you are working with a small set of data across a reasonably small range of values. This type of graph gives a clear view of the shape, any mode(s) and the range of a set of data. The numbers are already in order, so finding the median is fairly quick. And any outliers are quickly visible.

Ages of all of the Sales People at Stinky's Car Dealership.

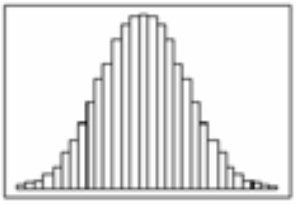


Describing a Numerical Distribution

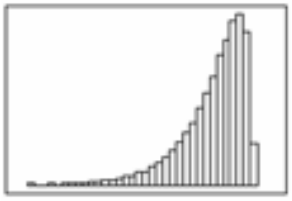
Shape

Once a graphical display is constructed, we can describe the distribution. When describing the distribution, we should be sure to address its shape. Although many graphs will not have a clear or exact shape, we can usually identify the shape as symmetrical or skewed. A **symmetrical** distribution will have a middle where we can draw an imaginary line through the center, and a fairly equal "look" on either side of that imaginary line. If you were to fold along the imaginary center line, the two sides would almost match up. Many symmetrical distributions are bell shaped, they will be tall in the middle with the two sides thinning

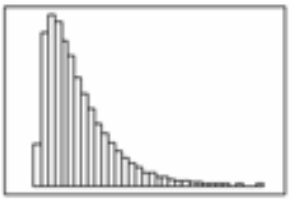
out. The sides are referred to as tails. A **skewed** distribution is one in which the bulk of the data is concentrated on one end, with the other side being a longer tail. The direction of the longer tail is the direction of the skew. Skewed right will have a longer tail to the right, or higher numbers. Skewed left will have a longer tail off to the left, or the lower values. Other shapes that you might see are uniform (almost consistent height all the way across) and bimodal (having two peaks in the distribution).



Symmetric
Bell shaped



Skewed to
the Left



Skewed to
the Right

Outliers

If there are any outliers, gaps, groupings, or other unusual features in the distribution, we should be sure to mention them. An **outlier** is a value that does not fit with the rest of the data. Some distributions will have several outliers, while others will not have any. We should always look for outliers because they can affect many of our statistics. Also, sometimes an outlier is actually an error that needs to be corrected. If you have ever 'bombed' one test in a class, you probably discovered that it had a big impact on your overall average in that class. This is because the mean will be affected by an outlier-it will be pulled toward it. This is another reason why we should be sure to look at the data, not just look at the statistics about the data. When an outlier is part of the data and we do not realize it, we can be misled by the mean to believe that the numbers are higher or lower than they really are.

Context

Do not forget that the graph, the numbers and the descriptions are all about something–its **context**. All of these elements of the distribution should be described in the context of the situation in question.

Center

The center of the distribution should always be included in the verbal analysis as well. People often wonder what the 'average is'. The measure for **center** can be reported as the median, the mean, or the mode. Even better, give more than one of these in your description. Remember that outliers effect the mean, but

do not effect the median. For example, the median of a list of data will stay in the center even when the largest value increases tremendously, but such a change would effect the mean quite a bit.

Spread

Another thing to include in the description is the spread of the data. The **spread** is the specific range of the data. When analyzing a distribution, we don't want to simply say that the range is equal to some number. It is much more informative to say that the data ranges from_____ to _____ (minimum value to maximum value). For example, if the news reports that the temperature in St. Paul had a range of 20° during a given week, this could mean very different temperatures depending on the time of year. It would be more informative to say something specific like, the temperature in St. Paul ranged from 68° to 88° last week.

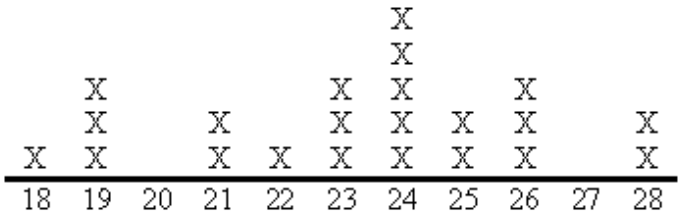
S.O.C.C.S.

So, when you describe the distribution of a numerical variable, there are several things to include. This text will use the acronym ***S.O.C.C.S!*** (shape, outliers, center, context, spread) to help us remember what characteristics to include in our descriptions.

Example 1

An anthropology instructor at the community college is interested in analyzing the age distribution of her students. The students in her Anthropology 102 class are: 21, 23, 25, 26, 25, 24, 26, 19, 18, 19, 26, 28, 24, 22, 24, 19, 23, 24, 24, 21, 23, and 28 years old. Organize the data in a dot plot. Calculate the mean, median, mode, and range for the distribution. Describe the distribution. Be sure to include the shape, outliers, center, context, and spread.

a) construct a dot plot



Ages of Students in Anthropology 102

b) mean- (18+19+19+19+21+21+22+23+23+23+24+24+24+24+24+25+25+26+26+26+28+28)/22 = 23.2727... mean = $\bar{x}$ = 23.27 **years old**

median- already listed in order, count to find "middle number", it is between 24 and 24, find mean of these two numbers (24+24)/2=24 median = Med = 24 years old

mode- look for most frequent age, it is 24 mode = 24 years old

range- min age is 18, max age is 28 range is 28 – 18 = 10 years or ages range from 18 to 28 years

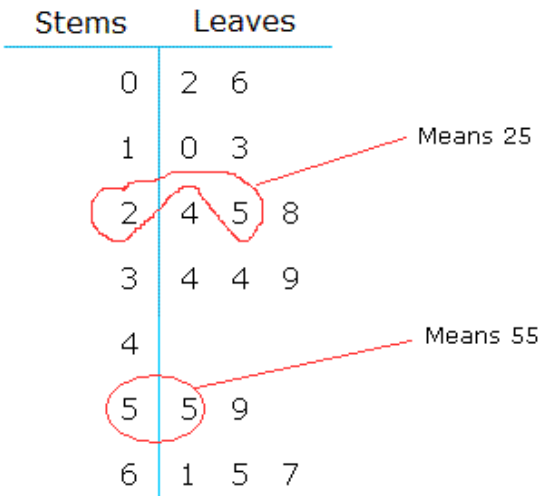
c) **describe-** address the shape, outliers, center, context, and spread of the distribution (This could be described as fairly symmetrical or slightly skewed to the left)

The distribution of student ages in this Anthropology 102 class is fairly symmetrical with no clear ouliers. The ages of students range from 18 to 28 years old. The median and mode for age are both 24 years old and the mean is 23.27 years. Thus, the typical student in this class is 23-24 years of age.

Stem Plots

In statistics, data is represented in tables, charts or graphs. One disadvantage of representing data in these ways is that the specific data values are often not retained. Using a stem plot is one way to ensure that the data values are kept intact. A **stem plot** is a method of organizing the data that includes sorting the data and graphing it at the same time. This type of graph uses the stem as the leading part of the data value and the leaf as the remaining part of the value. The result is a graph that displays the sorted data in groups or classes. A stem plot is used with numerical data when it will be helpful to see the actual values organized in order.

To construct a stem plot you must first determine the range of your distribution. Build the stems so that they cover the entire range, include every stem even if it will have no values after it. This will allow us to see the true shape of the distribution including outliers, whether it is skewed, and any gaps. Then place all of the "leaves" after the appropriate stems. Place the numbers in ascending order out and include all values, so repeats will show more than once. Some people like to put the numbers in order before they construct the stem plot, some like to try to put them in order as they make the plot, and others like to make a rough draft first without regard to order and then to make a final copy with the numbers in the correct order. Any of these methods will result in a correct stem plot.



Example 2

A researcher was studying the growth of a certain plant. She planted 25 seeds and kept watering, sunlight, and temperature as consistent as possible. The following numbers represent the growth (in centimeters)

of the plants after 28 days.

a) Construct a stem plot

b) Describe the distribution.

18	10	37	36	61
39	41	49	50	52
57	53	51	57	39
48	56	33	36	19
30	41	51	38	60

Solution

a) Construct a stem plot-the stem plot on the left was the first draft, the one on the right has the numbers in the correct order (ascending as you go out)

key 4/9 = 49cm

Plant height in CM Plant height in CM

Stem	Leaf
1	8, 0, 9
2	
3	7, 6, 9, 9, 3, 6, 0, 8
4	1, 9, 8, 1
5	0, 2, 7, 3, 1, 7, 6, 1
6	1, 0

Stem	Leaf
1	0, 8, 9
2	
3	0, 3, 6, 6, 7, 8, 9, 9
4	1, 1, 8, 9
5	0, 1, 1, 2, 3, 6, 7, 7
6	0, 1

b) Describe the distribution- Be sure to address shape, outliers, center, context, & spread.

The distribution of growth at 28 days ranged from 10 to 61 centimeters for these plants with the majority of plants growing at least 30cm. The median height was 41cm after 28 days. The shape is bimodal and there is a gap in the distribution because there are no plants in the 20-29 cm class. There are some possible low outliers, but no high outliers for plant growth.

Example 3

Sometimes a stem plot ends up looking too crowded. When the data is concentrated in a few rows, or 'classes', it can be difficult to determine what the shape is or whether there are any outliers in the data. In this example, the stem plot for the ages of a group of people was really concentrated in the 30s and 40s (plot on left). However, the statistician looking at this was not satisfied with the crowded appearance, so she decided to 'split' the stems. The resulting graph on the right, called a **split-stem plot**, shows very different results. Describe the distribution based on the split-stem plot.

STEM PLOT FOR AGES

2	0 2
3	3 3 4 5 5 6 7 7 7 8 9 9
4	0 1 4 4 5 7 7 7 7 8 8 8 9
5	1 2 2 4

SPLIT-STEM PLOT FOR AGES

2	0 2
2	
3	3 3 4
3	5 5 6 7 7 7 8 9 9
4	0 1 4 4
4	5 7 7 7 7 8 8 8 9
5	1 2 2 4

Solution

To split the stems, each stem was written twice. The top one is for the first half of the leaves in that class, and the second one is for the leaves in the second half of that class. For example the first stem of *4* gets 40 to 44, and the second *4* gets 45 to 49. So, when splitting stems into two, the number 5 is the cutoff for moving into the second part (just like rounding).

The split-stem plot shows that the distribution of ages in this example is bimodal and skewed to the left (lower numbers). It also shows that the ages of 20 and 22 appear to be low outliers. None of this was visible in the regular stem plot. Both plots show that the ages range from 20 to 54 years, with a median age of 41 years old and a mode age of 47 years old.

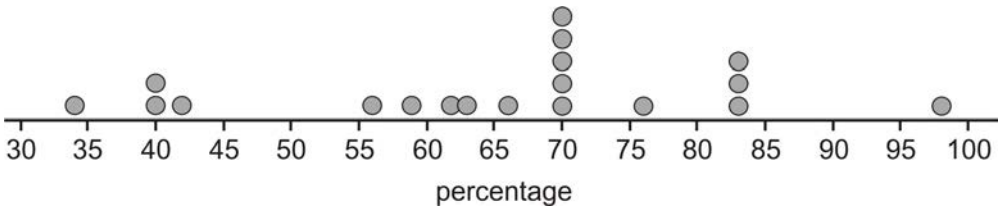
Section 5.3 Exercises

1) The following is data representing the percentage of paper packaging manufactured from recycled materials for a select group of countries.

Table 5.7: **Percentage of the paper packaging used in a country that is recycled. Source: National Geographic, January 2008. Volume 213 No.1, pg 86-87.**

Country	% of Paper Packaging Recycled
Estonia	34
New Zealand	40
Poland	40
Cyprus	42
Portugal	56
United States	59
Italy	62
Spain	63
Australia	66
Greece	70
Finland	70
Ireland	70
Netherlands	70
Sweden	76
France	76
Germany	83
Austria	83
Belgium	83
Japan	98

The dot plot for this data would look like this:



a) Calculate the mean, median, mode, and range for this set of data

b) Describe the distribution in context. Remember your S.O.C.C.S!

2) At the local veterinarian school, the number of animals treated each day over a period of 20 days was recorded.



28	34	23	35	16
17	47	05	60	26
39	35	47	35	38
35	55	47	54	48

a) Construct a stem plot for the data

b) Describe the distribution thoroughly. Remember your S.O.C.C.S!

3) The following table gives the percentages of municipal waste recycled by state in the United States, including the District of Columbia, in 1998. Data was not available for Idaho or Texas.

State	Percentage	State	Percentage	State	Percentage
AL	23	LA	14	OH	19
AK	7	ME	41	OK	12
AZ	18	MD	29	OR	28
AR	36	MA	33	PA	26
CA	30	MI	25	RI	23
CO	18	MN	42	SC	34
CT	23	MS	13	SD	42
DE	31	MO	33	TN	40
DC	8	MT	5	TX	NA
FL	40	NE	27	UT	19
GA	33	NV	15	VT	30
HI	25	NH	25	VA	35
ID	NA	NJ	45	WA	48
IL	28	NM	12	WV	20
IN	23	NY	39	WI	36
IA	32	NC	26	WY	5
KS	11	ND	21		
KY	28				

Source: http://www.zerowasteamerica.org

a) Create a split-stem plot for the data (there are 16 numbers in the 20-29 range)

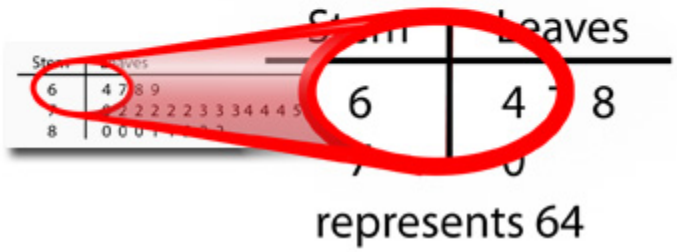
b) Find the median percentage for this data.

c) Which of the two plots you made gives a better view of the data? Explain.

d) Describe the distribution thoroughly. Remember your S.O.C.C.S! Specifically identify any states that stand out.

4) This stem plot is one that looks too crowded.

Stem	Leaves
6	4 7 8 9
7	0 2 2 2 2 2 3 3 3 4 4 4 5 5 6 6 6 7
8	0 0 0 1 1 2 2 2

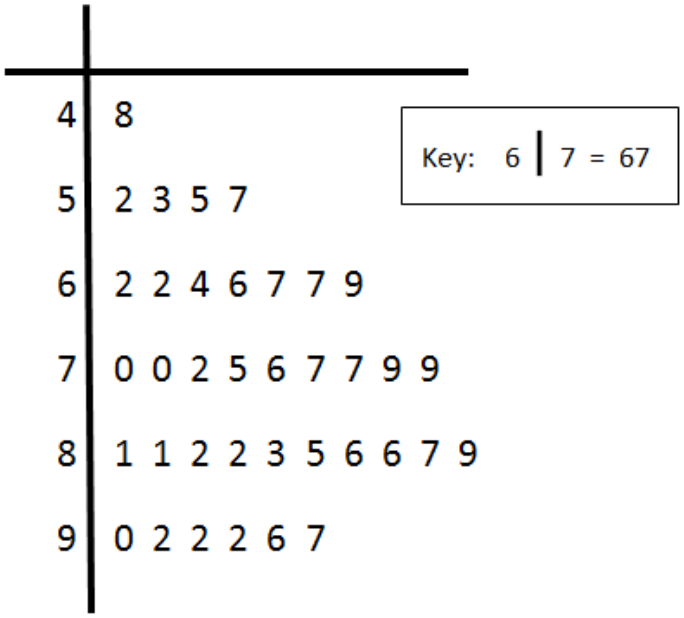


a) Create a split-stem plot for this example.

b) Name at least two things that are visible in the second plot that were not apparent in the first plot.

c) Invent a scenario that this data could represent.

5) Several game critics rated the ***Wow So Fit*** game, on a scale of 1 to 100 (100 being the highest rating). The results are presented in this stem plot:



a) Find the three measures of central tendency for the game rating data.

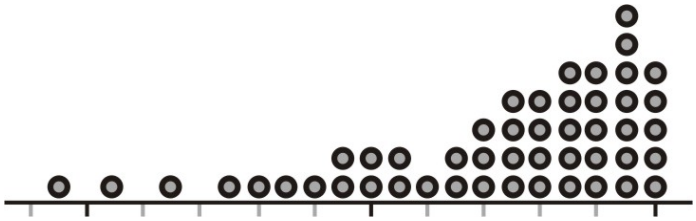
b) Which of these three measures of central tendency gives the best impression of the 'average' (typical) rating for this game? Explain.

6) These dot plots do not have any numbers or context. For each of the following dot plots:

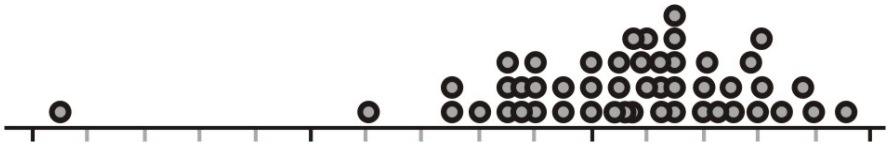
a) Identify the important features of the distribution (shape, outliers, center, spread).

b) Suggest a possible variable that might have such a distribution. (In other words, invent a context that fits the graph.)

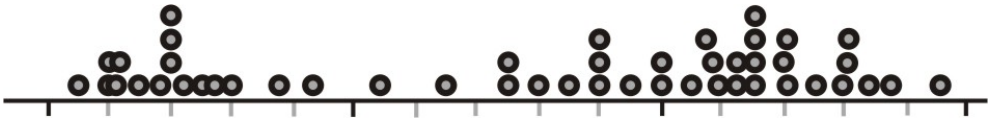
i)



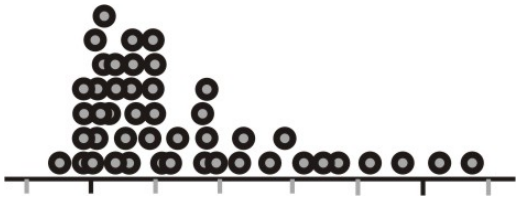
ii)



iii)



iv)



7) This table displays statistics for 21 of the Wild players for 2010-2011 regular season games. We are going to analyze the variable 'GP', which stands for games played.

Forwards & Defensemen

#	POS	PLAYER	GP	G	A	P	+/-	PIM	PP	SH	GW	S	S%
24	R	MARTIN HAVLAT	78	22	40	62	-10	52	3	0	4	229	9.6
9	C	MIKKO KOIVU	71	17	45	62	4	50	7	1	3	191	8.9
15	L	ANDREW BRUNETTE	82	18	28	46	-7	16	8	0	3	117	15.4
8	D	BRENT BURNS	80	17	29	46	-10	98	8	0	3	170	10.0
7	C	MATT CULLEN	78	12	27	39	-14	34	5	4	2	150	8.0
96	C	PIERRE-MARC BOUCHARD	59	12	26	38	-3	14	0	0	2	98	12.2
21	C	KYLE BRODZIAK	80	16	21	37	-4	56	2	1	1	126	12.7
20	R	ANTTI MIETTINEN	73	16	19	35	-3	38	8	0	4	168	9.5
22	R	CAL CLUTTERBUCK	76	19	15	34	-5	79	4	0	3	191	9.9
11	C	JOHN MADDEN	76	12	13	25	-9	10	1	1	4	107	11.2
3	D	MAREK ZIDLICKY	46	7	17	24	-6	30	3	0	0	53	13.2
55	D	NICK SCHULTZ	74	3	14	17	-4	38	0	0	0	46	6.5
12	R	CHUCK KOBASEW	63	9	7	16	-6	19	0	0	1	74	12.2
23	L	ERIC NYSTROM	82	4	8	12	-16	30	1	0	0	83	4.8
46	D	JARED SPURGEON	53	4	8	12	-1	2	2	0	1	38	10.5
4	D	CLAYTON STONER	57	2	7	9	5	96	0	0	1	40	5.0
16	R	BRAD STAUBITZ	71	4	5	9	-5	173	0	0	1	29	13.8
5	D	GREG ZANON	82	0	7	7	-5	48	0	0	0	55	0.0
19	C	PATRICK O'SULLIVAN	21	1	6	7	-1	2	0	0	0	37	2.7
48	L	GUILLAUME LATENDRESSE	11	3	3	6	2	8	1	0	1	18	16.7
25	D	CAM BARKER	52	1	4	5	-10	34	0	0	1	44	2.3

source: http://wild.nhl.com. July 25, 2011

a) Create a stem plot for the number of games played by these Wild players.

b) Calculate the mean, median, mode, range for the number of games played by these Wild players.

c) Describe the distribution of the number of games played by these players. Remember your S.O.C.C.S!

8) Now, you will examine the +/- data.

a) Find out what +/- stands for?

b) Construct a dot plot to show the +/- data.

b) Describe the distribution.

5.4 Numerical Data: Histograms



Learning Objectives

- Construct histograms
- Describe distributions including shape, outliers, center, context, and spread.

Histograms

When it is not necessary to show every value the way a stem plot would do, a histogram is a useful graph. Histograms organize numerical data into ranges, but do not show the actual values. The **histogram** is a summary graph showing how many of the data points falling within various ranges. Even though a histogram looks similar to a bar graph, it is not the same. Histograms are for numerical data and each 'bar' covers a range of values. Each of these 'bars' is called a **class** or **bin.** Histograms are a great way to see the shape of a distribution and can be used even when working with a large set of data.

The width of the bins is the most important decision when constructing a histogram. The bins need to be of consistent width (i.e. all cover a range of 10, or 25, etc.). It is generally a good idea to try to have 7 to 15 bins. Start with the range and divide by 10. This will give you a rough idea of how wide to make your bins. From there it becomes a judgment call as to what is a reasonable bin width. For example, it really does not make any sense to count by 11.24 just because that is what the range divided by 10 is equal to. In such a case, it might make more sense to count by 10's or 12's depending on the specific data.

Example 1

Suppose that the test scores of 27 students were recorded. The scores were: 8, 12,17, 22, 24, 28, 31, 37, 37, 39, 40, 42, 43, 47, 48, 51, 57, 58, 59, 60, 65, 65, 74, 75, 84, 88, 91. The lowest score was an 8 and the highest was a 91. Construct a histogram.

Solution

Plan bin width: The first step is to look at the range (91 - 8 = 83). Then divide the range by 10 (83/10 = 8.3). It doesn't make any sense to count by bins of 8.3 points, so we may use 8, or 10, or 12. Next we look at where to start. The first number is 8. It doesn't make any sense to start counting at 8 either, or to end at 91. We will probably want to start from 0 and end at 100, counting by 10's should work nicely.

**Where to begin, and what to count by are not obvious to a calculator or many computer software programs. The graphing calculator would probably start at 8, and count by 8.3. Leaving you with bins of [8 -16.3); [16.3-24.6); [24.6 -32.9); etc. So, if you are using technology to create a histogram, you will generally need to fix the window so that the bins make sense.*

Mark horizontal axis: Mark your scale along the horizontal axis to cover your entire range and to count by your decided upon bin width. Include numbers.

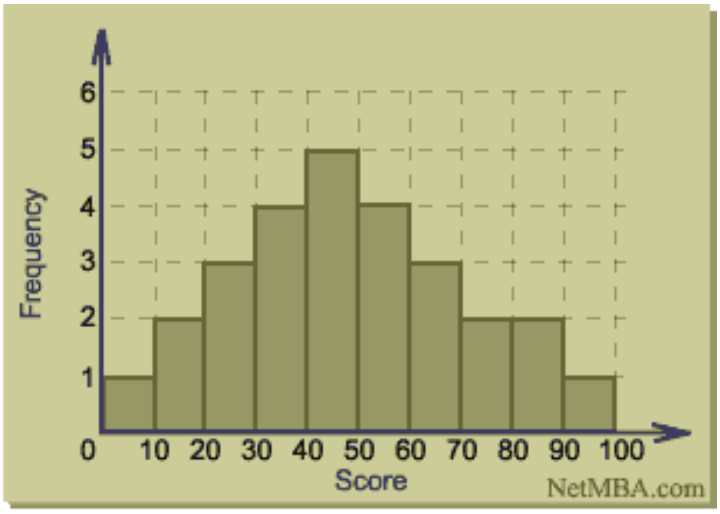
Count number of values within each bin: How many values falls between 0 and <10? One, so we make the bin one unit tall. Between 10 and <20? Two, so we make the bin two units tall, etc. A frequency table may be helpful here. You need to know how tall to make each bin. You especially need to know how tall to make the tallest of the bins.

Mark vertical axis: Your vertical axis needs to reach the height of the tallest bin. Mark your vertical axis by consistent steps so that it will reach the number needed. Include numbers.

**For instance, if you need to get to 2,460; then you should probably count by steps of 250's or even a larger number.*

Make your histogram: Make the bins the correct heights, shade or color them in, add labels including any units, a title, and a key if needed.

TEST SCORES



Test score histogram. http://www.netmba.com

The bins in this example are [0 to 10); [10 to 20); etc. This means that zero up to, but not including, 10 are in the first bin (9.999 would be in bin #1, but 10 would be in bin #2).

You may be creating your histograms with paper and pencil. However, the graphing calculators are a great way to create histograms as well. It takes a little practice to learn how to adjust the windows, but you have the opportunity to try out different bin widths without needing to erase or start all over. Also, you may want to see how to create histograms in excel. When you use a graphing calculator to create your graphs, you should sketch what the calculator shows you. Your sketch should look similar to the graphing window shown, and will still need labels and titles.



Example 2

a) Construct a histogram to look at the distribution of acceptance rates for these U.S. Universities.

b) Describe your findings.

The following table gives a list of the acceptance rate for applicants to twelve U.S. universities.
(Source: *Time Almanac 2004*)

College or Univeristy	Percent Accepted
Harvard University	11
Yale University	16
Princeton University	12
Johns Hopkins University	32
New York University	29
M. I. T.	16
Duke University	26
Carnegie Mellon University	36
George Washington University	49
Northwestern University	33
American University	72
Cornell University	31

http://jcsites.juniata.edu

Solution

a) Try this on your calculator: Enter the data in a list and set up a histogram.

Plan bin width: Determine the range (72 -11= 61). Divide by 10 (61/10 = 6.1) to get a rough idea of a good bin width. We can use a variety of bin width of 5, 7.5, 8, or 10, etc. We must start before the minimum of 11 (start at 0 or 10), and pass the maximum of 72 (80).

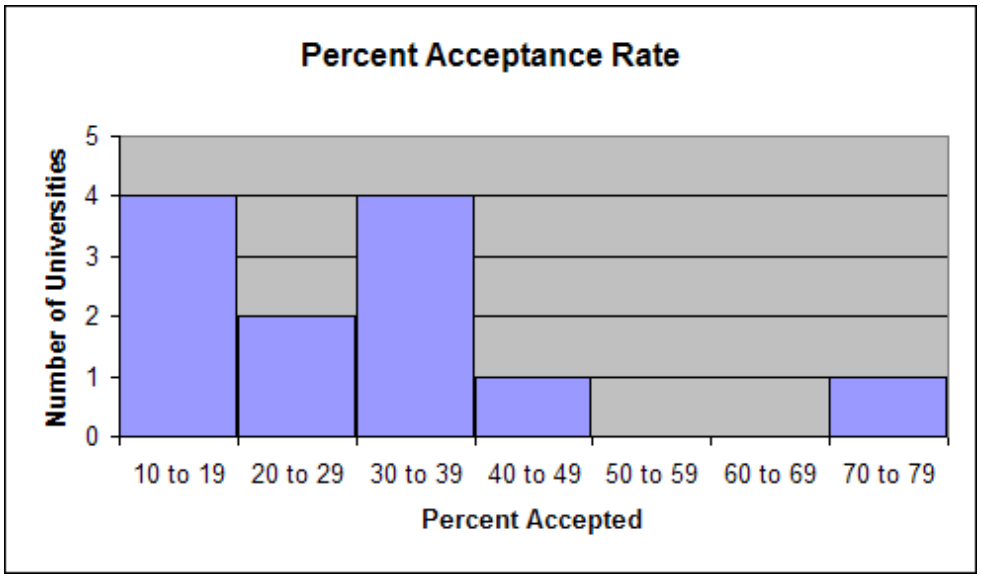
After trying a few of these bins, we decide to use bins of 10, starting at 10 and ending at 80. Here is the window that was used: {x-min =10, x-max=80, x-scl=10, y-min=-2, y-max=5, y-scl=1}

Mark horizontal axis: Mark your scale along the horizontal axis to cover your entire range and to count by your decided upon bin width. Include numbers.

Count number of values within each bin: A frequency table may be helpful here. You need to know how tall to make each bin. You especially need to know how tall to make the tallest of the bins.

Mark vertical axis: Your vertical axis needs to reach the height of the tallest bin. Mark your vertical axis by consistent steps so that it will reach the number needed. Include numbers.

Make your histogram: Make the bins the correct heights, shade or color them in, add labels including and units, a title, and a key if needed.

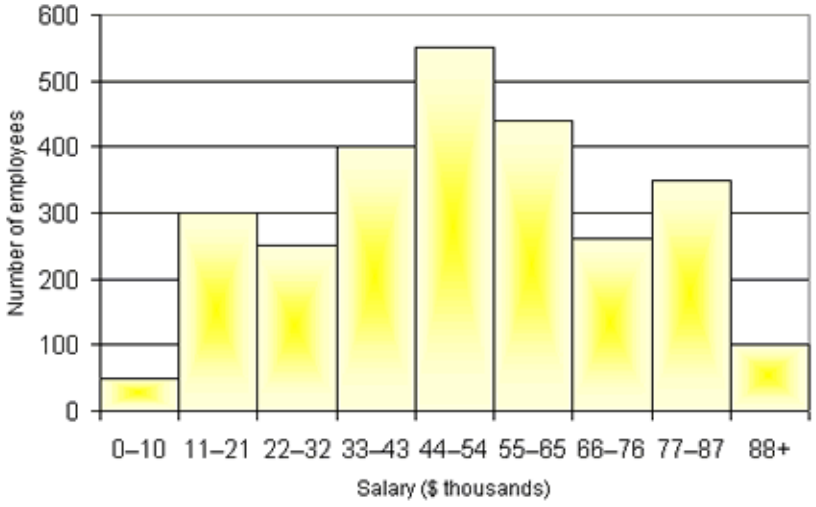


b) Describe: The median and mean are difficult to identify from just a histogram. You will often only be able to estimate them. In this case, we were given all of the original data so we can find the exact values. When possible, identify outliers specifically.

The median acceptance rate for these Universities is 30%. The percent of students applying, who are accepted to these universities ranged from 11% to 72%. However, the 72% was an extremely high outlier because the next highest rate was 49%. The majority of these schools accepted 36% or fewer of those who applied. The distribution is heavily skewed to the right because of the high outlier of American University.

Section 5.4 Exercises

1) This graph shows the distribution of salaries (in thousands of dollars) for the employees of a large school district. Answer the questions that follow.



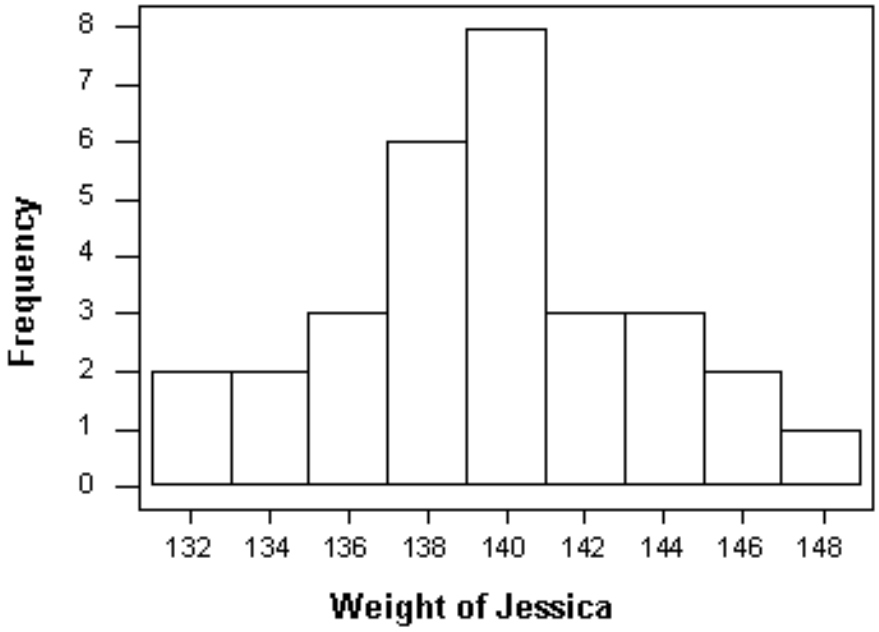
Source: http://4.bp.blogspot.com

a) Approximately how many employees make $77,000 or more per year?

b) What is the bin width here? Be careful.

c) Without calculating anything, how would you describe the typical salary of an employee of this school district?

2) Jessica is a freshman at the University of Minnesota, Duluth. She has been watching her weight because she is afraid of gaining that 'freshman fifteen' she keeps hearing about. She has weighed herself every Monday morning since school started. Here is a histogram showing the results in pounds of all of these *Monday-Morning-Weigh-In's*.



a) Describe the distribution. Remember your S.O.C.C.S!

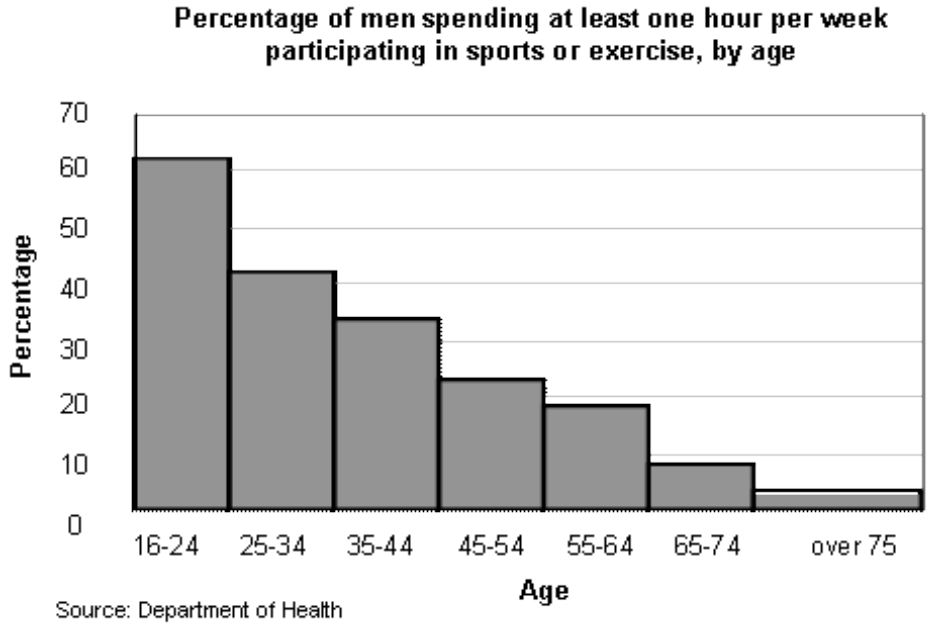
b) What is the range for the bin that has 6 observations?

c) For her height, Jessica feels that 140 lbs. is her ideal weight. What percent of the time has she been within 5 lbs. of her ideal weight?

3) Pretend you are a journalist.

a) What do you notice that is wrong with this graph?

b) Based on only what you can see in the graph and labels, write several sentences that could go with this graph. (Think S.O.C.C.S!) Ignore the mistake from part (a).



Men and exercise graph: http://www2.le.ac.uk

4) Here are the statistics from several of the Minnesota Wild players. We are going to analyze the Penalties in Minutes (PIM) data.

Forwards & Defensemen

#	POS	PLAYER	GP	G	A	P	+/-	PIM	PP	SH	GW	S	S%
24	R	MARTIN HAVLAT	78	22	40	62	-10	52	3	0	4	229	9.6
9	C	MIKKO KOIVU	71	17	45	62	4	50	7	1	3	191	8.9
15	L	ANDREW BRUNETTE	82	18	28	46	-7	16	8	0	3	117	15.4
8	D	BRENT BURNS	80	17	29	46	-10	98	8	0	3	170	10.0
7	C	MATT CULLEN	78	12	27	39	-14	34	5	4	2	150	8.0
96	C	PIERRE-MARC BOUCHARD	59	12	26	38	-3	14	0	0	2	98	12.2
21	C	KYLE BRODZIAK	80	16	21	37	-4	56	2	1	1	126	12.7
20	R	ANTTI MIETTINEN	73	16	19	35	-3	38	8	0	4	168	9.5
22	R	CAL CLUTTERBUCK	76	19	15	34	-5	79	4	0	3	191	9.9
11	C	JOHN MADDEN	76	12	13	25	-9	10	1	1	4	107	11.2
3	D	MAREK ZIDLICKY	46	7	17	24	-6	30	3	0	0	53	13.2
55	D	NICK SCHULTZ	74	3	14	17	-4	38	0	0	0	46	6.5
12	R	CHUCK KOBASEW	63	9	7	16	-6	19	0	0	1	74	12.2
23	L	ERIC NYSTROM	82	4	8	12	-16	30	1	0	0	83	4.8
46	D	JARED SPURGEON	53	4	8	12	-1	2	2	0	1	38	10.5
4	D	CLAYTON STONER	57	2	7	9	5	96	0	0	1	40	5.0
16	R	BRAD STAUBITZ	71	4	5	9	-5	173	0	0	1	29	13.8
5	D	GREG ZANON	82	0	7	7	-5	48	0	0	0	55	0.0
19	C	PATRICK O'SULLIVAN	21	1	6	7	-1	2	0	0	0	37	2.7
48	L	GUILLAUME LATENDRESSE	11	3	3	6	2	8	1	0	1	18	16.7
25	D	CAM BARKER	52	1	4	5	-10	34	0	0	1	44	2.3

a) Construct a histogram for the penalties in minutes for the Wild players included on that list.

b) Describe the distribution. Remember your S.O.C.C.S!

5) Refer to the Age at Inauguration of the United States Presidents data in this table.

George Washington	57	Abraham Lincoln	52	Herbert Hoover	54
John Adams	61	Andrew Johnson	56	Franklin Roosevelt	51
Thomas Jefferson	57	Ulysses S. Grant	46	Harry S. Truman	60
James Madison	57	Rutherford B. Hayes	54	Dwight D. Eisenhower	62
James Monroe	58	James A. Garfield	49	John F. Kennedy	43
John Quincy Adams	57	Chester A. Arthur	50	Lyndon B. Johnson	55
Andrew Jackson	61	Grover Cleveland	47	Richard M. Nixon	56
Martin Van Buren	54	Benjamin Harrison	55	Gerald Ford	61
William Henry Harrison	68	Grover Cleveland	55	Jimmy Carter	52
John Tyler	51	William McKinley	54	Ronald Reagan	69
James K. Polk	49	Theodore Roosevelt	42	George H. W. Bush	64
Zachary Taylor	64	William Howard Taft	51	Bill Clinton	46
Millard Fillmore	50	Woodrow Wilson	56	George W. Bush	54
Franklin Pierce	48	Warren G. Harding	55	Barack Obama	47
James Buchanan	65	Calvin Coolidge	51		

Source: http://qrc.depaul.edu

a) Construct a histogram for the distribution of ages of the U.S. Presidents at their inauguration.

b) Calculate the mean, median, mode.

c) Which of these three measures of central tendency best describe the typical inauguration age? Explain.

6) Sketch a histogram that fits the following scenarios:

a) Symmetrical with a few high outliers and a few low outliers.

b) Strongly skewed right with no outliers.

c) Bimodal and symmetrical.

d) Skewed left with a few outliers.

e) Doesn't fit any of the descriptions we have learned.

5.5 Numerical Data: Box Plots & Outliers



Learning Objectives

- Calculate the five number summary for a set of numerical data
- Construct box plots
- Calculate IQR and standard deviation for a set of numerical data
- Determine which numerical summary is more appropriate for a given distribution
- Determine whether or not any values are outliers based on the 1.5*(IQR) criterion
- Describe distributions in context– including shape, outliers, center, and spread

Box Plots

A box plot (also called box-and-whisker plot) is another type of graph used to display data. A **box plot** divides a set of numerical data into quarters. It shows how the data are dispersed around a median, but does not show specific values in the data. It does not show a distribution in as much detail as does a stem plot or a histogram, but it clearly shows where the data is located. This type of graph is often used when the number of data values is large or when two or more data sets are being compared. The center and spread of the distribution are very obvious from the graph. It is easy to see the range of the values as well as how these values are distributed around the middle value. The smaller the box, the more consistent the data values are with the median of the data. The shape of the box plot will give you a general idea of the shape of the distribution, but a histogram or stem plot will I do this more accurately. Any outliers will show up as long whiskers.The box in the box plot contains the middle 50% of the data, and each 'whisker' contains 25% of the data.

The Five Number Summary

In order to divide into fourths, it is necessary to find five numbers. This list of five values is called the **five number summary**. The numbers in the list are {minimum value, Quartile 1, Median, Quartile 3, maximum value}. We have already learned how to find the median of a set of numbers (put in order and find the middle value), and the minimum and maximum are the smallest and largest numbers. Now we will learn how to find the quartiles.

$$5\#\ \text{sum} = \{\min, Q_1, \text{Med}, Q_3, \max\}$$

Quartiles

The first step is to list all of the numbers in order from least to greatest. The minimum and maximum are now on the ends of the list and we can count in to find the median–circle these three values. Finding the quartiles is just like finding the median. **Quartile 1** is the 'median' of all of the values to the left of the median (do NOT include the median itself). **Quartile 3** is the 'median' of all of the values to the right of the median (do not include the median).

Constructing a Box Plot

Now list the five number summary in order {min, Q1, Med, Q3, max). The next step is to mark an axis that covers the entire range of the data. Mark the numbers along the axis before you make the box plot, so that the resulting plot shows the shape of the data. The last step is to place a dot above the axis for the 5 numbers from the five number summary, and then to make a 'box' through the second and fourth dots, mark a line through the middle dot to show the median, and mark 'whiskers' from the box out to the first and fifth dots.



Example 1

You have a summer job working at Paddy's Pond which is a recreational fishing spot where children can go to catch salmon which have been raised in a nearby fish hatchery and then transferred into the pond. The cost of fishing depends upon the length of the fish caught ($0.75 per inch). Your job is to transfer 15 fish into the pond three times a day. But, before the fish are transferred, you must measure the length of each one and record the results. Below are the lengths (in inches) of the first 15 fish you transferred to the pond. Calculate the five number summary, and construct a box plot for the lengths of these fish.

Length of Fish (in.)				
13	14	6	9	10
21	17	15	15	7
10	13	13	8	11

Solution

Since box plots are based on the median and quartiles, the first step is to organize the data in order from smallest to largest.

6	7	8	9	10
10	11	13	13	13
14	15	15	17	21

$$6, 7, 8, 9, 10, 10, 11, \boxed{13}, 13, 13, 14, 15, 15, 17, 21$$

The minimum is the smallest number (min = 6), and the maximum is the largest number (max = 21). Next, we need to find the median. This has an odd number of data, so the median of all the data is the value in the middle position (Med = 13). There are 7 numbers before and 7 numbers after 13. The next step is the find the median of the first half of the data – the 7 numbers before the median, but not including the median. This is called the lower quartile since it marks the point above the first quarter of the data. On the graphing calculator this value is referred to as Q_1.

$$6, 7, 8, \boxed{9}, 10, 10, 11$$

Quartile 1 is the median of the lower half of the data ($Q_1 = 9$).

This step must be repeated for the upper half of the data – the 7 numbers above the median of 13. This is called the upper quartile since it is the point that marks the third quarter of the data. On the graphing calculator this value is referred to as Q_3.

$$13, 13, 14, \boxed{15}, 15, 17, 21$$

Quartile 3 is the median of the upper half of the data ($Q_3 = 15$).

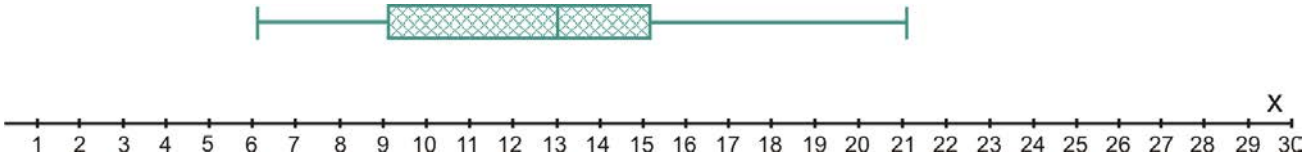
Now that the five numbers have all been determined, it is time to construct the actual graph. The graph is drawn above a number line that includes all the values in the data set (graph paper works very well since the numbers can be placed evenly using the lines of the graph paper). For this examle we will need to mark from at least 6 to at least 21. Be sure to mark your axis before you start to construct the box plot. Next, represent the following values by placing dots above their corresponding values on the number line:

Minimum – 6	Quartile 1 – 9	Median – 13
Quartile 3 – 15	Maximum – 21	

The five data values listed above are often called the five number summary for the data set and are necessary to graph every box plot.

Make the 'box' part around the Q_1 and Q_3 values, make 'whiskers' out to the min and max values, and make a vertical line to show the location of the median. This will complete the box plot.

Length of fish (in inches) 5# summary = {6, 9, 13, 15, 21}



The five numbers divide the data into four equal parts. In other words:

- One-quarter of the data values are located between 6 and 9
- One-quarter of the data values are located between 9 and 13
- One-quarter of the data values are located between 13 and 15
- One-quarter of the data values are located between 15 and 21

.

More Measures of Spread

Range

We have already learned how to find the range of a set of data. The range represents the entire spread of all of the data.

The formula for calculating the range is:

max - min = range

Inner Quartile Range

The quartiles give us one more measure of spread called the inner quartile range. The **inner quartile range (IQR)** is the range between the lower and upper quartile. To find the IQR, subtract the quartile 1 value from the quartile 3 value (Q_3 - Q_1 = IQR). The IQR represents the spread, or range, of the middle 50% of the data. The IQR is a measure of spread that is used when the median is the measure of central tendency.

The formula for calculating the IQR is:

Q_3 - Q_1 = IQR

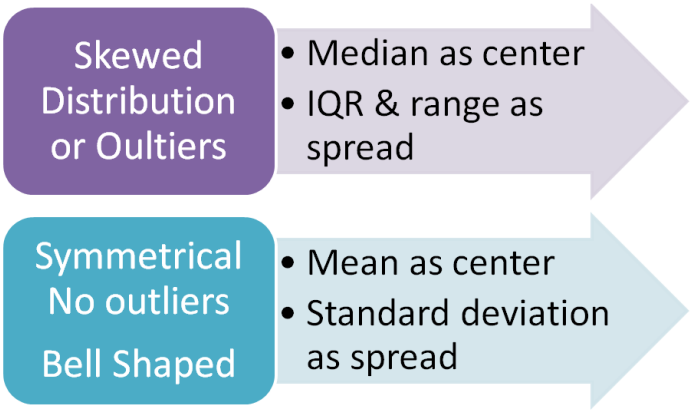
Standard Deviation

Another measure of spread that is used in statistics is called the standard deviation. The **standard deviation** measures the spread around the mean. This value is more difficult to calculate than range or IQR, but the formula used takes all of the data values in the distribution into account. Standard deviation is the appropriate measure of spread when the mean is the measure of center. However, the standard deviation is easily affected by outliers or skewness because every value is calculated in the formula. The symbol for standard deviation of a sample is s (on the graphing calculators it is $S_{x)}$ and for a population it is σ (sigma).

The standard deviation can be any number zero or greater. It will only be equal to zero if there is no spread (i.e. all values are exactly the same). The more spread out the data is, the larger the standard deviation will be. The standard deviation is most appropriate when you have a very symmetrical, bell-shaped distribution called a normal distribution. We will study this type of distribution in chapter 7.

Which Numerical Summary Should We Use?

We have learned several statistics that are measures of central tendency and several that are measures of spread. How do we know which one(s) to use? The mean and standard deviation go together. And, the median will go with the IQR (or range). The most important thing to remember is that the mean and the standard deviation are both affected by outliers and by skewness in a distribution. So if either of these is present, then the mean and standard deviation are not appropriate. However, it is always an option, and often interesting to calculate all of the statistics and compare them to one another. The general guidelines are:


Skewed Distribution or Oultiers
• Median as center
• IQR & range as spread
Symmetrical No outliers
Bell Shaped
• Mean as center
• Standard deviation as spread

How to Calculate the Standard Deviation

In order to calculate the standard deviation you must have all of the values. Then you follow these steps:

1. Calculate the mean of the values.
2. Subtract the mean from each data value. These are the individual deviations.
3. Each of these deviations is squared.
4. All of the squared deviations are added up.
5. This total of the squared deviations is divided by one less than the number of deviations. This is the variance.
6. Take the square root of the variance. This is the standard deviation.

The formula for calculating the variance is:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x - \bar{x})^2$$

The formula for calculating standard deviation is:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x - \bar{x})^2}$$

As you can probably tell, this formula is very time consuming when you have a large set of data. Also, it is easy to make a mistake in your calculations. We will show the process with a small set of data, but generally we will use our calculator to find the standard deviation.

Example 2

Find the mean and standard deviation for these numbers:{12, 15, 14, 17, 19}

Solution

1. Calculate the mean of the values. $\frac{(12+15+14+17+19)}{5} = 15.4$
2. Subtract the mean from each data value. These are the individual deviations.
3. Each of these deviations is squared.
4. All of the squared deviations are added up.
5. This total of the squared deviations is divided by one less than the number of deviations. This is the variance.
6. Take the square root of the variance. This is the standard deviation.

Data values	Value – mean = deviation	Deviation squared
x	$(x - \bar{x})$	$(x - \bar{x})^2$
12	$(12 - 15.4) = -3.4$	$(-3.4)^2 = 11.56$
15	$(15 - 15.4) = -0.4$	$(-0.4)^2 = 0.16$
14	$(14 - 15.4) = -1.4$	$(-1.4)^2 = 1.96$
17	$(17 - 15.4) = 1.6$	$(1.6)^2 = 2.56$
19	$(19 - 15.4) = 3.6$	$(3.6)^2 = 12.96$
Sum of the squared deviations		29.2
Variance = $\frac{sum}{n-1}$		$s^2 = \frac{29.2}{5-1} = 7.3$
Standard Deviation = $\sqrt{s^2}$		$s = \sqrt{7.3} = 2.7019$

The mean of these numbers is 15.4 and the standard deviation is 2.7019.

The standard deviation is tedious to calculate. For any problem where you are asked to calculate the standard deviation, you may use your calculator or a computer to find it.

Example 3

After one month of growing, the heights of 30 parsley seed plants were measured and recorded. The measurements (in inches) are shown in the table below.

Table 5.8: **Heights of Parsley (in.)**

22	28	30	40	38	18
11	37	12	34	49	17
25	37	46	39	8	27
16	38	18	23	26	14
6	26	23	33	11	26

a) Calculate the five number summary and construct a box plot to represent the data.

b) Describe the distribution.

c) Calculate the mean and standard deviation.

d) Calculate the median, and IQR

Solution

a) five number summary and box plot:

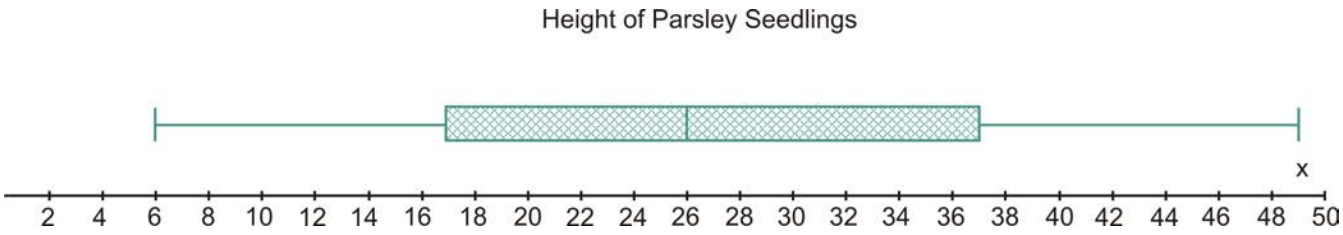
order the values– The data organized from smallest to largest is shown in the table below. (You could use your calculator to quickly sort these values)

Table 5.9: **Heights of Parsley (in.)**

6	8	11	11	12	14
16	17	18	18	22	23
23	25	26	26	26	27
28	30	33	34	37	37
38	38	39	40	46	49

5# summary– This time there is an even number of data values so the median will be the mean of the two middle values. $Med = \frac{26+26}{2} = 26$ (We will not use the median, but we do use the values on either side of it when finding quartiles). The median of the lower half is the number in the 8th position which is 17. The median of the upper half is the number in the 22nd position (or 8th from the top) which is 37. The smallest number is 6 and the largest number is 49.

5# summary = {6, 17, 26, 37, 49}



b) describe–don't forget your S.O.C.C.S!

The heights of these parsley plants ranged from 6 inches to 49 inches after one month. The distribution is very symmetrical and does not contain any outliers. The median height for these parsley plants was 26 inches tall. The middle 50% of the plants were all between 17 inches and 37 inches tall.

c) The mean and standard deviation were calculated using the TI-84+.

$\bar{x} = 25.9333$

$s = 11.4709$

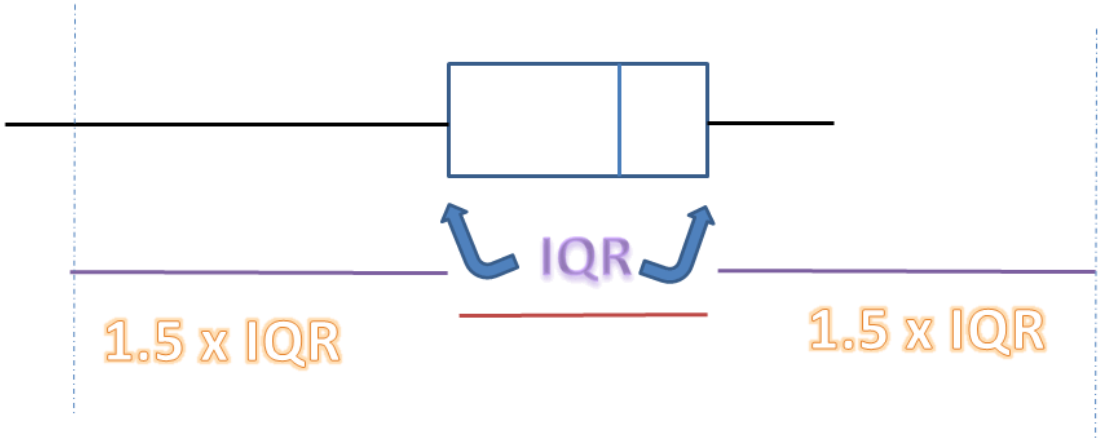
d) The median is part of the five number summary. The **IQR** = Q_3 - Q_1 = 37 - 17 = 20

$Med = 26$

$IQR = 20$

Outliers

We have been noticing some values that appear to be outliers, but have not defined a specific distance to be considered an outlier. The common **outlier test**, used to determine whether or not any of the values are outliers uses the IQR. This outlier test, often called the 1.5*(IQR) Criterion, says that any value that is more than one and one-half times the width of the IQR box away from the box is an outlier.



Any value past the cutoff points (shown as dashed lines above) will be considered an outlier.

The above example would have at least one low outlier, but no high outliers.

Cutoff value for LOW OUTLIERS:

$Q_1 - 1.5^*(IQR)$ **any value less than this number is considered a low outlier*

Cutoff value for HIGH OUTLIERS:

$Q_3 + 1.5^*(IQR)$ **any value greater than this number is considered a high outlier*

Example 4

Test the sodium in the McDonald's® sandwiches listed in the table in exercise number one for outliers. Use the 1.5*(IQR) Criterion. Show your steps.

Solution

Calculate the five number summary for the Amount of Sodium (in mg) $fivenumbersummary = \{520, 835, 1095, 1285, 2070\}$

First find the IQR: $IQR = 1285 - 835 = 450$

Test for low outliers: $Q1 - 1.5(IQR)$

$835 - 1.5(450) = 160$

Test for high outliers: $Q3 + 1.5(IQR)$

$1285 + 1.5(450) = 1960$

Check the data to see if we have any outliers:

We have no sandwiches with less than 160 mg sodium, so we have no low outliers.

We have one value that is greater than this cutoff of 1960 mg. The Angus Bacon & Cheese burger has 2070 mg of sodium, so we have one high outlier.

Section 5.5 Exercises

1) Here is some nutritional information about a few of the sandwiches on the McDonald's® menu.

Nutrition Facts	Serving Size	Calories	Calories from Fat	Total Fat (g)	% Daily Value**	Saturated Fat (g)	% Daily Value**	*Trans* Fat (g)	Cholesterol (mg)	% Daily Value**	Sodium (mg)	% Daily Value**
Sandwiches												
Hamburger	3.5 oz (100 g)	250	80	9	13	3.5	16	0.5	25	9	520	22
Cheeseburger	4 oz (114 g)	300	110	12	19	6	28	0.5	40	13	750	31
Double Cheeseburger	5.8 oz (165 g)	440	210	23	35	11	54	1.5	80	26	1150	48
McDouble	5.3 oz (151 g)	390	170	19	29	8	42	1	65	22	920	38
Quarter Pounder® with Cheese+	7 oz (198 g)	510	230	26	40	12	61	1.5	90	31	1190	50
Double Quarter Pounder® with Cheese++	9.8 oz (279 g)	740	380	42	65	19	95	2.5	155	52	1380	57
Big Mac®	7.5 oz (214 g)	540	260	29	45	10	50	1.5	75	25	1040	43
Big N' Tasty®	7.2 oz (206 g)	460	220	24	37	8	42	1.5	70	23	720	30
Big N' Tasty® with Cheese	7.7 oz (220 g)	510	250	28	43	11	54	1.5	85	28	960	40
Angus Bacon & Cheese	10.2 oz (291 g)	790	350	39	60	17	87	2	145	49	2070	86
Angus Deluxe	11.1 oz (314 g)	750	350	39	60	16	82	2	135	45	1700	71
Angus Mushroom & Swiss	10 oz (283 g)	770	360	40	61	17	85	2	135	46	1170	49

Source: http://nutrition.mcdonalds.com. July 27, 2011.

Determine the median and the IQR for the following data regarding the McDonald's® sandwiches:

a) Calories from fat

b) Cholesterol

2) Analyze the calories for these McDonald's® sandwiches.

a) Calculate the five number summary and construct an accurate box plot for the calories for these sandwiches.

b) Use the outlier test to determine whether there are any outliers for calories. Test for both high and low outliers. Show your steps.

c) Describe the distribution in context- Remember your S.O.C.C.S!

3) Analyze the sodium content further.

a) Construct a box plot for sodium.

b) Calculate the median and IQR for sodium (see example 4).

c) Calculate the mean and standard deviation for sodium (use a calculator).

Now **remove the high outlier** from the data.

d) Re-calculate the median and IQR for sodium with the Angus Bacon & Cheese data removed. Did either value change from part (b)?

e) Re-calculate the mean and standard deviation for sodium with the Angus Bacon & Cheese data removed. Did either value change from part (c)?

4) The following table shows the potential energy that could be saved by manufacturing each type of material using the maximum percentage of recycled materials, as opposed to using all new materials.

Table 5.10:

Manufactured Material	**Energy Saved (millions of BTU's per ton)**
Aluminum Cans	206
Copper Wire	83
Steel Cans	20
LDPE Plastics (e.g. trash bags)	56
PET Plastics (e.g. beverage bottles)	53
HDPE Plastics (e.g. household cleaner bottles)	51
Personal Computers	43
Carpet	106
Glass	2
Corrugated Cardboard	15
Newspaper	16
Phone Books	11
Magazines	11
Office Paper	10

Source: National Geographic, January 2008. Volume 213 No., pg 82-

a) Calculate the five number summary and construct an accurate box plot for the Energy Saved data.

b) Use the outlier test to determine whether there are any outliers. Show your steps.

c) Calculate the mean and standard deviation for the Energy Saved data. How do the mean and the median compare?

d) Delete any outliers. Recalculate the five number summary, mean and standard deviation. Which values changed?

5) The Burj Dubai will become the world's tallest building when it is completed. It will be twice the height of the Empire State Building in New York. The chart lists the 15 tallest buildings in the world (as of 12/2007).

Table 5.11:

Building	**City**	**Height (ft)**
Taipei 101	Tapei	1671
Shanghai World Financial Center	Shanghai	1614
Petronas Tower	Kuala Lumpur	1483
Sears Tower	Chicago	1451
Jin Mao Tower	Shanghai	1380
Two International Finance Center	Hong Kong	1362
CITIC Plaza	Guangzhou	1283
Shun Hing Square	Shenzen	1260
Empire State Building	New York	1250
Central Plaza	Hong Kong	1227
Bank of China Tower	Hong Kong	1205
Bank of America Tower	New York	1200
Emirates Office Tower	Dubai	1163
Tuntex Sky Tower	Kaohsiung	1140
Burj Dubai		

a) Calculate the five number summary for these 15 buildings and construct an accurate box plot.

b) Use the outlier test to determine whether there are any outliers among these 15 buildings. Test for both high and low outliers. Show your steps.

c) Describe the shape of the distribution. Remember your S.O.C.C.S!

d) Within what range of heights are the middle 50% of these buildings?

6) The table shows the U. S. Presidents' ages at inauguration.

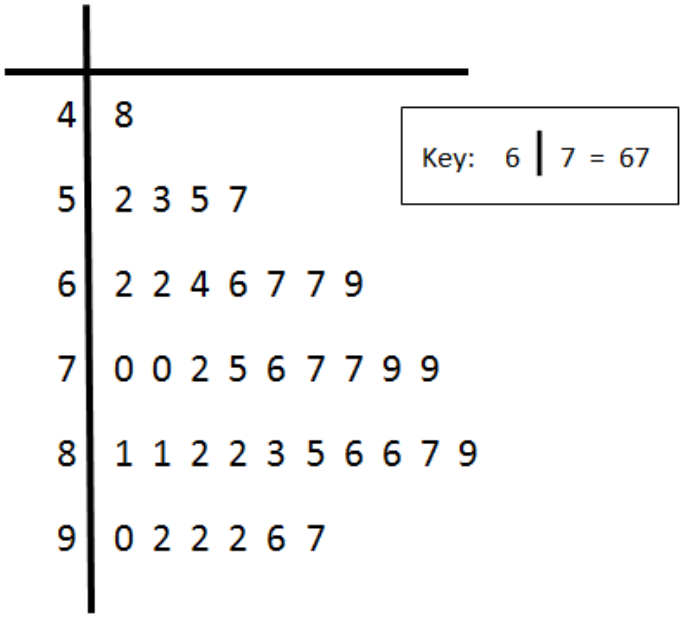
George Washington	57	Abraham Lincoln	52	Herbert Hoover	54
John Adams	61	Andrew Johnson	56	Franklin Roosevelt	51
Thomas Jefferson	57	Ulysses S. Grant	46	Harry S. Truman	60
James Madison	57	Rutherford B. Hayes	54	Dwight D. Eisenhower	62
James Monroe	58	James A. Garfield	49	John F. Kennedy	43
John Quincy Adams	57	Chester A. Arthur	50	Lyndon B. Johnson	55
Andrew Jackson	61	Grover Cleveland	47	Richard M. Nixon	56
Martin Van Buren	54	Benjamin Harrison	55	Gerald Ford	61
William Henry Harrison	68	Grover Cleveland	55	Jimmy Carter	52
John Tyler	51	William McKinley	54	Ronald Reagan	69
James K. Polk	49	Theodore Roosevelt	42	George H. W. Bush	64
Zachary Taylor	64	William Howard Taft	51	Bill Clinton	46
Millard Fillmore	50	Woodrow Wilson	56	George W. Bush	54
Franklin Pierce	48	Warren G. Harding	55	Barack Obama	47
James Buchanan	65	Calvin Coolidge	51		

a) Construct a box plot for the Presidents ages at inauguration.

b) Make a statement, in context, about what the 'box' part of the box plot tells you.

c) Describe the distribution. Remember your S.O.C.C.S! Identify any unusual values specifically.

7) Several game critics rated the ***Wow So Fit*** game, on a scale of 1 to 100 (100 being the highest). The results are presented in this stem plot:



a) Calculate the five number summary for the ***Wow So Fit*** data.

b) Construct a box plot for the data.

c) Describe this distribution.

d) Calculate the mean and standard deviation.

5.6 Numerical Data: Comparing Data Sets

Learning Objectives

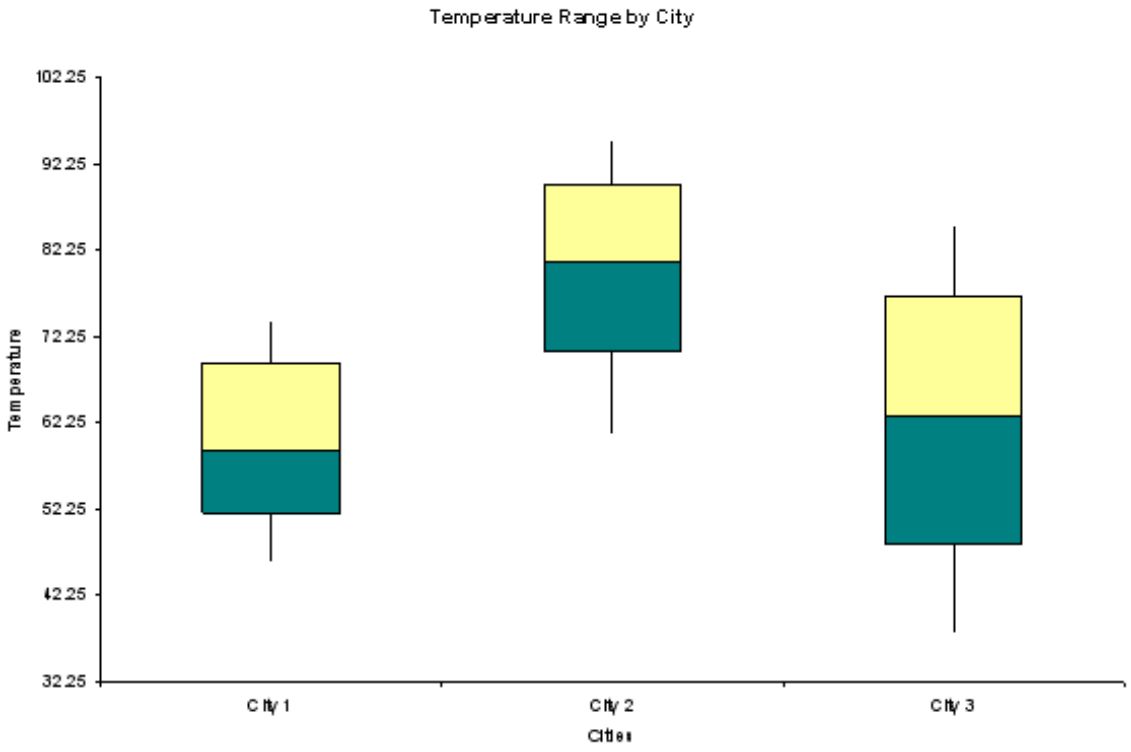
- Construct parallel box plots
- Construct back-to-back stem plots
- Compare more than one set of numerical data in context

Parallel Box Plots

Parallel box plots (also called side-by-side box plots) are very useful when two or more numerical data sets need to be compared. The graphs of the **parallel box plots** are plotted, one parallel to the other, along the same number line. This can be done vertically or horizontally and for as many data sets as needed.

Example 1

The figure shows the distributions of the temperatures for three different cities. By graphing the three box plots along the same axis, it becomes very easy to compare the temperatures of the three cities. What are some conclusions that can be drawn about the temperatures in these three cities?



http://www.mathworksheetscenter.com

Solution

Here are some conclusions, based on the graphs, that might be made. Think S.O.C.C.S! And, be sure to compare the distributions to one another, using statistics to support your observations.

- ***Quartile 1 for City 2 is higher than the quartile 3 in City 1 and the median in City 3. Also, the minimum temperature in City 2 is at about the median for the other two cities.***
- ***City 2 is generally warmer than both of the other cities. Cities 1 and 3 have nearly the same median temperature, around 60° to 63°. Whereas, the median temperature in City 2 is around 82°.***
- ***City 3 has a much larger range in temperatures (35° to 85°), than City 1 (45° to 75°) or City 2 (62° to 95°). Thus, the temperature in City 1 is the most consistent of the three.***
- ***The temperature distributions in all three cities are fairly symmetrical and none have any outliers.***

Comparing Numerical Data Sets

When you are given numerical sets of data for more than one variable and asked to compare them, it will be necessary to construct graphical representations for each data set. In order to compare them to one another the scales must match. When comparing more than one box plot, we construct parallel box plots. When using histograms, we can match the horizontal and vertical scales so that the separate histograms can 'line up'. Dot plots will work the same way as histograms. Such comparisons are also possible when working with stem plots. Two sets of numerical data can simply share the stems in the middle, with one set's 'leaves' going to the right and the other set's 'leaves' going to the left. On both sides of the plot, the 'leaves' will go in numerical order out. Plots like these are called **back-to-back stem plots**.

Once you have constructed any of these types of comparative graphical representations (on the same scale,) you can make observations about how the data sets are the same and how they are different. Just as we have been doing up to this point, those comparisons should be done in context. The observations made might address the shapes of the distributions and whether or not any outliers are present. It is important to compare the centers of the distributions (means, medians, or modes). And, the spreads of the distributions should also be addressed (ranges, IQRs, or standard deviations).

Example 2

A teacher gave the same physics exam to her two sections of physics. She has been wondering whether the first period and fifth period classes are learning the same amount as one another. She constructed this back-to-back stem plot to compare the test scores for the two different classes.

a) Calculate the five number summary for both classes.

b) Calculate the mean and standard deviation for both classes.

c) Compare the two classes' test scores in context.

Class A Leaves	Stems	Class B Leaves
8 0	6	0 0
5 0	7	0 1 3 3 5 6 7
6 4	8	4 5 6
6 4 4 2 1 0	9	1 2
0 0	10	

http://www.basic-mathematics.com

Solution

a) The numbers in the stem plots are already in order, so these statistics could be found by hand or with a graphing calculator.

Five number summary for Class A $\{60, 75, 90.5, 94, 100\}$

Five number summary for Class B $\{60, 71, 75.5, 85, 92\}$

b) These statistics are most efficiently found using a graphing calculator.

Class A mean $\bar{x} \approx 85.7143$ Class A standard deviation $s \approx 12.6396$

Class B mean $\bar{x} \approx 76.6429$ Class B standard deviations $s \approx 10.0507$

C) Comparison

Overall, Class A did better on this test than Class B did. Class A's scores on this test are skewed to the left, but Class B's scores are skewed to the right. Neither class has any outliers among the test scores. Class A has a mean score of about 9 points higher (85.7 compared to 76.6) and a median score of 15 points higher (90.5 compared to 75.5). The overall range for the two classes is fairly similar, but the Class A students' scores were less consistent. The ranges (32 and 40), IQRs (14 and 19), and standard deviations (10.1 and 12.6), all show that Class B's scores are less spread out than Class A's scores.

Example 3

An oil company claims that its premium grade gasoline contains an additive that significantly increases gas mileage. They conducted the following experiment in an effort to prove their claim. They selected 15 drivers who all drove the same make, model and year of car. Starting with an empty gas tank, each car was filled with 45L of one of the two types of gasoline (selected in a random order). The driver was asked to drive until the fuel light warning came on. The number of kilometers was recorded and then the car was filled with the other type of gasoline (whichever they had not used yet). The process was repeated and the number of kilometers was again recorded. The results below show the number of kilometers each car traveled.

Regular Gasoline						Premium Gasoline				
640	570	640	580	610		659	619	639	629	664
540	555	588	615	570		635	709	637	633	618
550	590	585	587	591		694	638	689	589	500

Display each set of data to explain whether or not the claim made by the oil company is true or false.

Solution

order the data–list the values in order for each set of data

Regular Gasoline						Premium Gasoline				
540	550	555	570	570		500	589	618	619	629
580	585	587	588	590		633	635	637	638	639
591	610	615	640	660		659	664	689	694	709

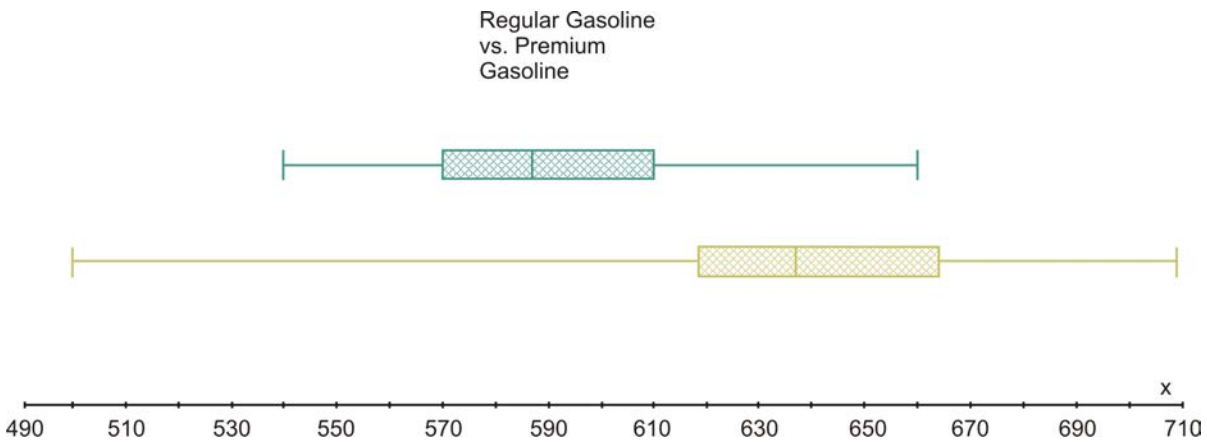
5 # summaries- Determine the five number summary for each set of data separately. Be sure to report your five number summary, whether asked to or not.

Five-Number Summary				
		Regular Gasoline		Premium Gasoline
Smallest #		540		500
Q_1		570		619
Median		587		637
Q_3		610		664
Largest #		660		709

box plots –Mark your number axis so that it covers the entire range needed – smallest minimum to largest maximum (we need 500 to 709 for these two data sets). Then graph each box plot along the same axis, but parallel to each other. This allows for the two data sets to be easily compared to one another.

Key: blue = regular gasoline

gold = premium gasoline

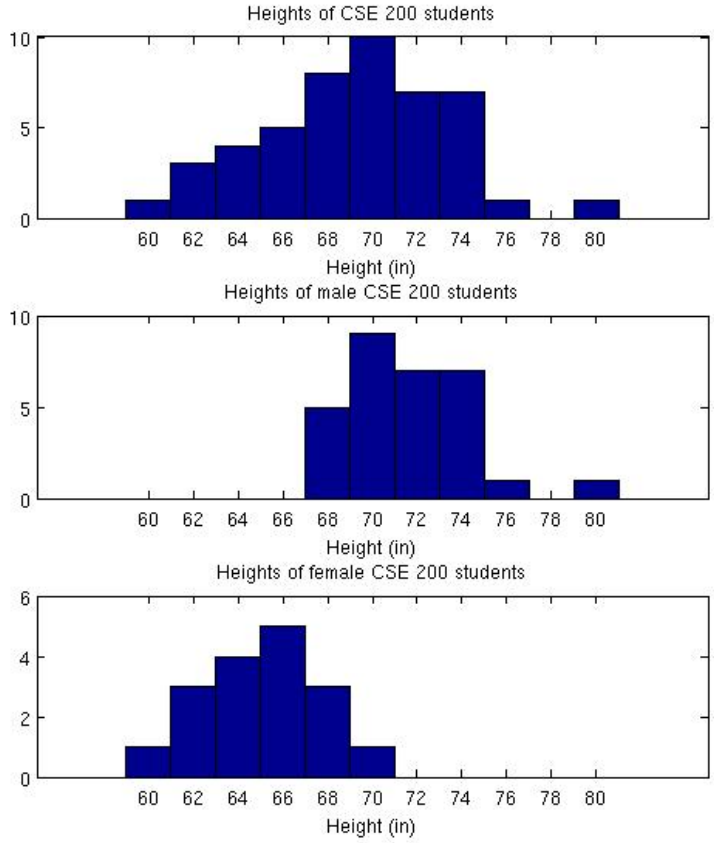


conclusions– make comparisons by looking for any similarities and differences between the two distributions. Remember your S.O.C.C.S!

Based on this experiment, the number of kilometers that the cars were able to travel on the premium gasoline was greater than the number of kilometers that the same cars were able to travel with the regular gasoline. The median number of kilometers for premium gasoline was 637, compared to 587 for regular gas. The first quartile for premium was higher than the third quartile for regular. Also, 25% of those with the premium gasoline went further than all of those using regular gasoline. The distribution for the regular fuel is slightly skewed to the right, but doesn't have any outliers. However the premium distribution is strongly skewed to the left toward one outlier on the low end (500 km). Based on these results, it appears that the additive in the premium gasoline does improve gas mileage for this make and model of car. Further tests should be done on other types of vehicles.

Example 4

The heights of a group of students are all included in the first histogram. The second histogram only contains the data from the male students and the third is a graph of the heights of only the girls. Explain what this histograms show.



Solution

The range of heights of all students in this group is approximately 20 inches. However, the female heights only range about 11 inches and the male heights only range about 13 inches. The females' height distribution is the most symmetrical of all three. There is one male whose height is a high outlier, but none for the females. The median height for the class is around 70 inches, for males it is slightly higher around 72 inches, and for females it is around 65 inches tall. In general, the female students tend to be shorter than the male students.

Section 5.6 Exercises

1) Compare the %Daily Value for Total Fat(g) to the %Daily Value for Saturated Fat(g) for these McDonald's® sandwiches.

a) Calculate the five number summary for both %Daily Values.

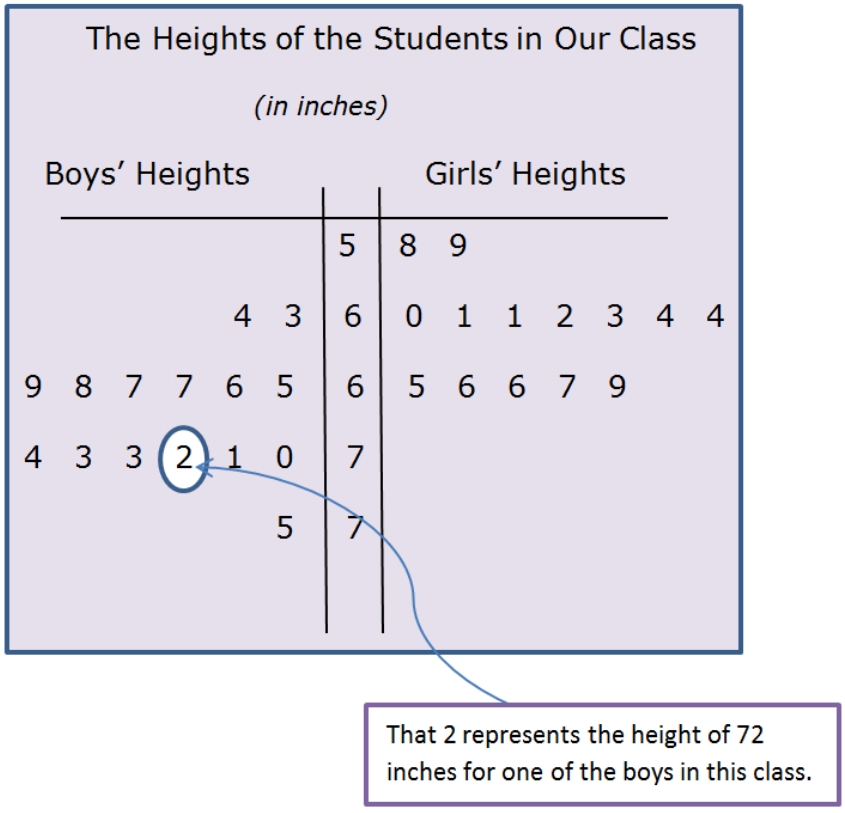
b) Construct parallel box plots for both.

d) Make at least four observations to compare these two distributions.

Nutrition Facts	Serving Size	Calories	Calories from Fat	Total Fat (g)	% Daily Value**	Saturated Fat (g)	% Daily Value**	*Trans* Fat (g)	Cholesterol (mg)	% Daily Value**	Sodium (mg)	% Daily Value**
Sandwiches												
Hamburger	3.5 oz (100 g)	250	80	9	13	3.5	16	0.5	25	9	520	22
Cheeseburger	4 oz (114 g)	300	110	12	19	6	28	0.5	40	13	750	31
Double Cheeseburger	5.8 oz (165 g)	440	210	23	35	11	54	1.5	80	26	1150	48
McDouble	5.3 oz (151 g)	390	170	19	29	8	42	1	65	22	920	38
Quarter Pounder® with Cheese+	7 oz (198 g)	510	230	26	40	12	61	1.5	90	31	1190	50
Double Quarter Pounder® with Cheese++	9.8 oz (279 g)	740	380	42	65	19	95	2.5	155	52	1380	57
Big Mac®	7.5 oz (214 g)	540	260	29	45	10	50	1.5	75	25	1040	43
Big N' Tasty®	7.2 oz (206 g)	460	220	24	37	8	42	1.5	70	23	720	30
Big N' Tasty® with Cheese	7.7 oz (220 g)	510	250	28	43	11	54	1.5	85	28	960	40
Angus Bacon & Cheese	10.2 oz (291 g)	790	350	39	60	17	87	2	145	49	2070	86
Angus Deluxe	11.1 oz (314 g)	750	350	39	60	16	82	2	135	45	1700	71
Angus Mushroom & Swiss	10 oz (283 g)	770	360	40	61	17	85	2	135	46	1170	49

Source: http://nutrition.mcdonalds.com. July 27, 2011.

2) The heights of the students in a statistics class were all measured to the nearest inch. The results are presented in this back-to-back stem plot. Notice that it is also a split stem plot. The girls' heights are ordered out to the right on the right side. And the boys' heights are ordered out to the left on the left side.


The Heights of the Students in Our Class
(in inches)
Boys' Heights
Girls' Heights
5 | 8 9
4 3 | 6 | 0 1 1 2 3 4 4
9 8 7 7 6 5 | 6 | 5 6 6 7 9
4 3 3 2 1 0 | 7
5 | 7
That 2 represents the height of 72 inches for one of the boys in this class.

a) Compute the standard deviation, the range, and the IQR for both girls and boys.

b) Compare the spread for the two groups, based on your answers to (a), in context.

c) Compute the mean, median, and mode for both boys and girls.

d) Compare the center for the two groups, based on your answers to (c), in context.

e) Compare the shape of the distributions, based on the graph, in context.

3) Compare the results of the Probability and Statistics District Common Assessment for two statistics classes.

CLASS 3:

45, 37, 14, 42, 24, 33, 41, 16, 39, 24, 38, 35, 35, 32, 51, 46, 30, 42, 25, 37, 37, 19, 26, 23, 28, 38, 16, 35, 21

CLASS 4:

35, 37, 25, 44, 31, 27, 26, 35, 24, 41, 30, 29, 30, 29, 40, 25, 38, 31, 42, 46, 37, 32, 20, 40, 35, 29, 25, 31, 27, 43, 27, 30, 38, 36, 37

a) Construct back-to-back stem plots (use split-stems) for these two classes.

b) Calculate the five number summaries for both classes.

c) Calculate the following statistics for both classes: mean, standard deviation, mode, range, and IQR.

d) Compare and contrast the two distributions. This should be in context and you should make at least four distinct observations.

4) The number of home-runs during a season is one of the statistics recorded about baseball players. The following table has the number of home-runs (over many seasons) for several of the best hitters in baseball. Compare the home-run hitting performance of these exceptional baseball players.

- **Babe Ruth: 54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22**
- **Mark McGwire: 49, 32, 33, 39, 22, 42, 9, 9, 39, 52, 58, 70, 65, 32, 29**
- **Barry Bonds: 16, 25, 24, 19, 33, 25, 34, 46, 37, 33, 42, 40, 37, 34, 49, 73, 46**
- **Roger Maris: 13, 23, 26, 16, 33, 61, 28, 39, 14, 8**

a) Construct Parallel Box Plots for the four players. Be sure to use the same scale for all four graphs and to label each graph.

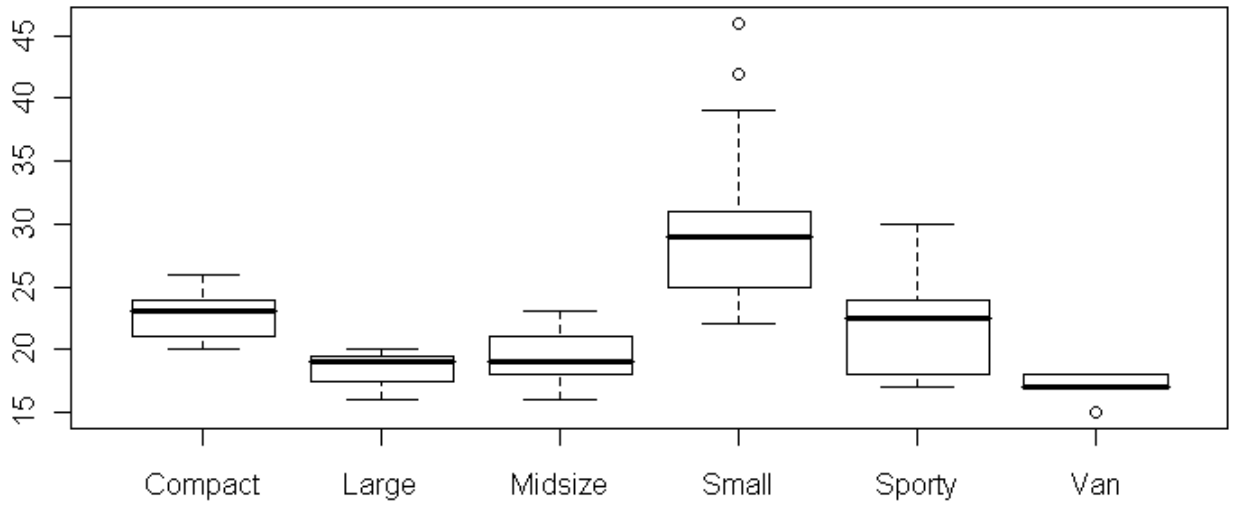
b) Calculate the following for all four players: $\bar{x}$ = _____ s_x= _____ IQR = _____

5 # summary = {___, ___, _____, _____, _____}

c) Test for outliers, for all four players, using the 1.5*IQR criterion-(*show work*).

d) Compare and contrast the four distributions. This should be in context and you should make at least four distinct observations.

5) The following box plots show the average miles per gallon (city) for various types of vehicles. Comment on what these parallel box plots show. This should be in context and include at least 4 distinct observations.

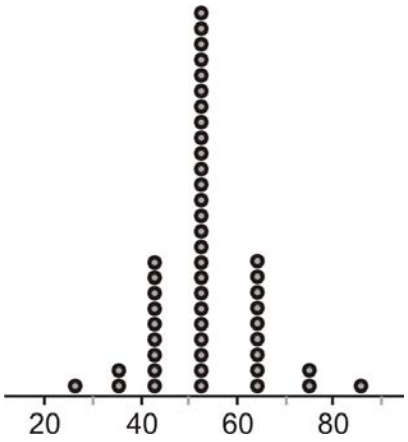


boxplot(MPG.city~Type) # base package

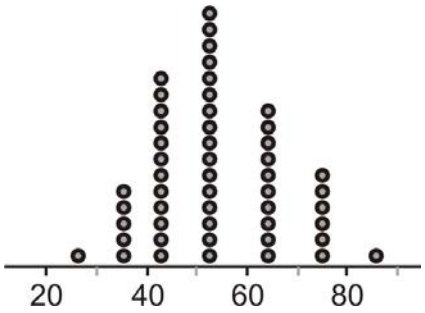
http://www.fort.usgs.gov

6) Refer to the four dot plots to answer the questions that follow.

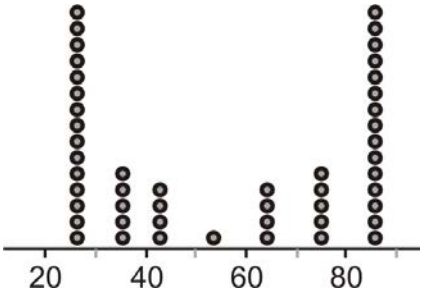
Graph I



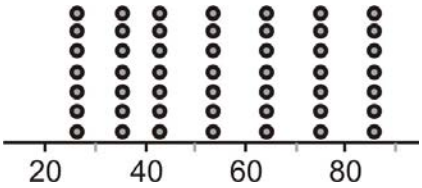
Graph II



Graph III



Graph IV



a) Identify the overall shape of each distribution.

b) How would you characterize the center(s) of these distributions?

c) Name at least two statistics that would most likely be the same for all four of these distributions.

d) Which of these distributions has the smallest standard deviation? Which of these distributions has the largest standard deviation? Explain.

e) For which of these distributions would it be appropriate to use the mean and standard deviation as numerical summaries? For which would the five number summary be more appropriate?

5.7 Chapter 5 Review

Chapter 5 Summary

In this chapter, we have learned that when working with a set of data it is important to choose an appropriate type of graphical display so that we can see what the data looks like. Bar graphs and pie charts are useful ways to display categorical data. Time plots are line graphs that help us to see how a given variable has changed over a specified period of time. And, when working with numerical data, we have learned how to make dot plots, stem plots, histograms, and box plots. It is also possible to make graphs so that comparisons can be made between more than one data set. Back-to-back stem plots and parallel box plots are two such types of graphs.

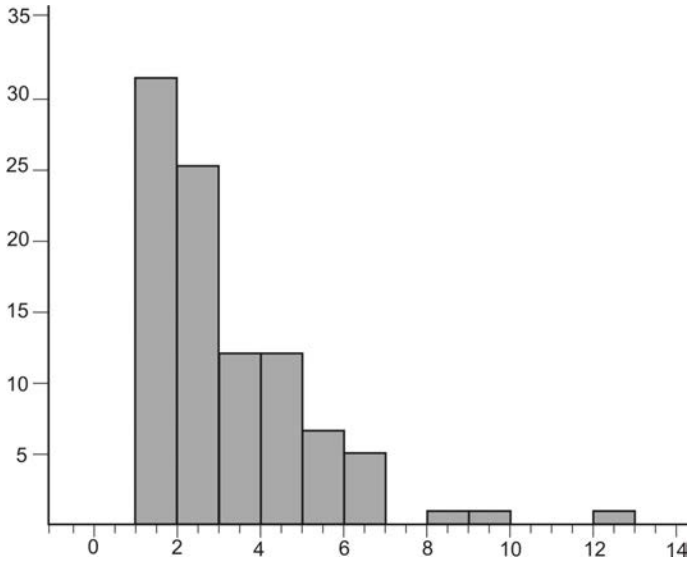
The next step is to analyze the data set(s) by calculating numerical statistics. The statistics that give us an idea of where the center of the data is are the mean, median, and mode. These statistics are measures of central tendency, and give us an idea of where the 'average' of the data lies. The range, inner quartile range (IQR), and the standard deviation are all measures of the spread of a set of data. We have also learned how to calculate the five number summary, which divides a set of data into quarters and allows us to construct a box plot.

Once the graphs are constructed and the statistics are calculated, we have learned to describe what these show. When describing a numerical set of data, in addition to explaining where the center and spread are, we also describe the shape of the distribution and whether or not any outliers are present in the data. The shapes that we focused on are symmetrical distributions and skewed distributions, remembering that the direction of the skew is toward the tail or outliers. We learned to make appropriate conclusions and comparisons that are based on the data, graphs, and statistics. Statisticians should avoid opinions and judgment statements as much as possible.

We learned that the 1.5*(IQR) Criterion can be used to determine whether or not any data values are outliers. And, that the mean and standard deviation are easily changed, so these statistics are not the appropriate measures of center and spread when working with data that contains outliers or is skewed.

Chapter 5 Review Exercises

1) Multiple-Choice: Which of the following can be inferred from this histogram?



a) The mode is 1.

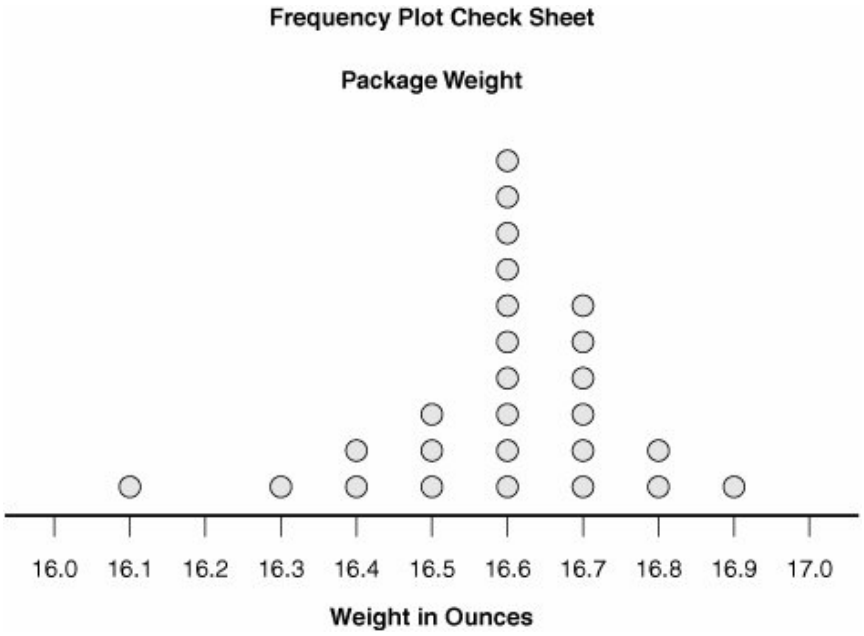
b) mean < median

c) median < mean

d) The distribution is skewed left.

e) None of the above can be inferred from this histogram.

2) The owner of a small company is trying to determine whether he should go with a different company for his shipping needs. He needs to analyze the weights of the packages that his company ships out. This graph shows the distribution of the weight of packages that were shipped during the last month.



a) Calculate the mean, standard deviation, mode, and range for this data. Use a calculator for mean and standard deviation.

b) Determine the five number summary for this data. Construct a box plot for this data.

c) Which of these two graphs is more informative? Explain.

d) He figures that he will save money with the new company on any packages that weigh less than 16.75 ounces. What percent of packages weighed more than 16.75 ounces?

3) After some bullying issues were brought to light in a big high school, a committee was formed to study the issue. A questionnaire was designed that contained several questions related to bullying and safety. A stratified random sample was selected that included students from all four grade levels. The table that follows shows the responses to one of the questions on the questionnaire.

Student Responses to the Question:

"I Feel Safe at School"

n = sample size	Strongly Disagree	Disagree	Agree	Strongly Agree
1003	133	274	529	67

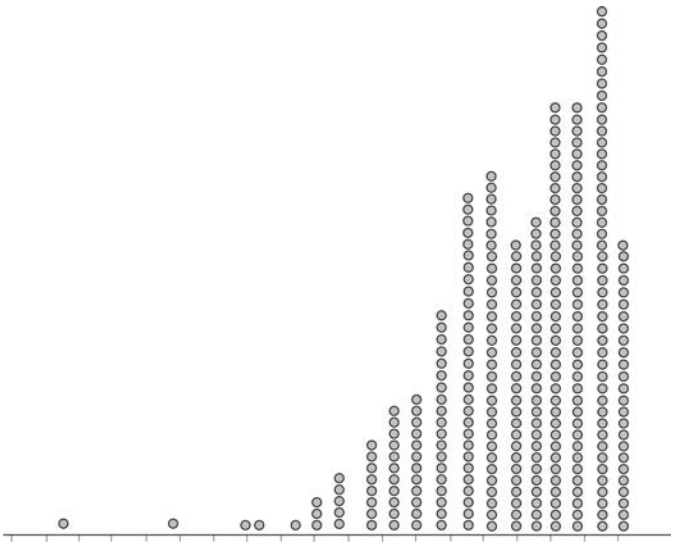
a) Use Excel or Google Docs to create a pie chart that shows the results of this survey question. Be sure to include labels, percents, a title, and a key if needed.

b) Describe what the graph shows in context. Be sure to include percents to support your observations.

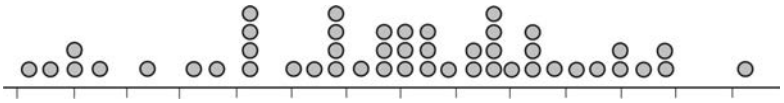
c) Comment on whether or not the committee should be concerned. Explain.

In questions 4-7, match the distribution with the choice of the correct real-world situation that best fits the graph.

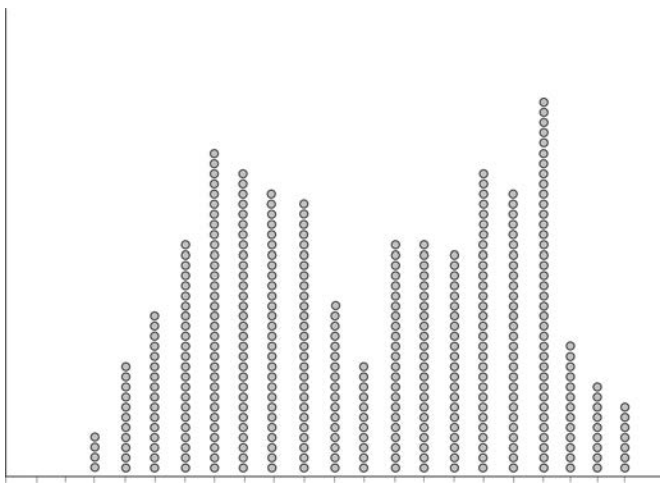
4)



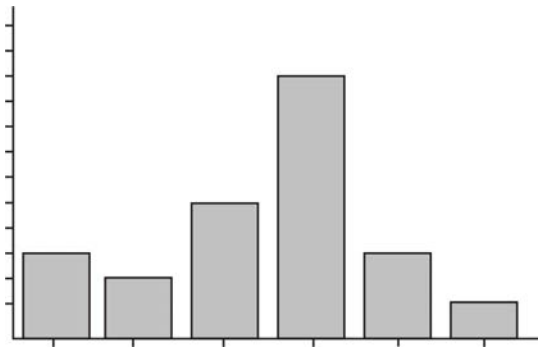
5)



6)



7)



a) Andy collected and graphed the heights of all the 12th grade students in his high school.

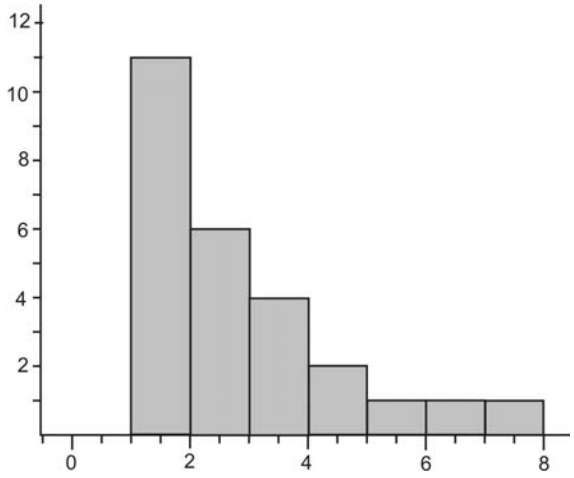
b) Brittany asked each of the students in her statistics class to bring in 20 pennies selected at random from their pocket or piggy bank. She created a plot of the dates of the pennies.

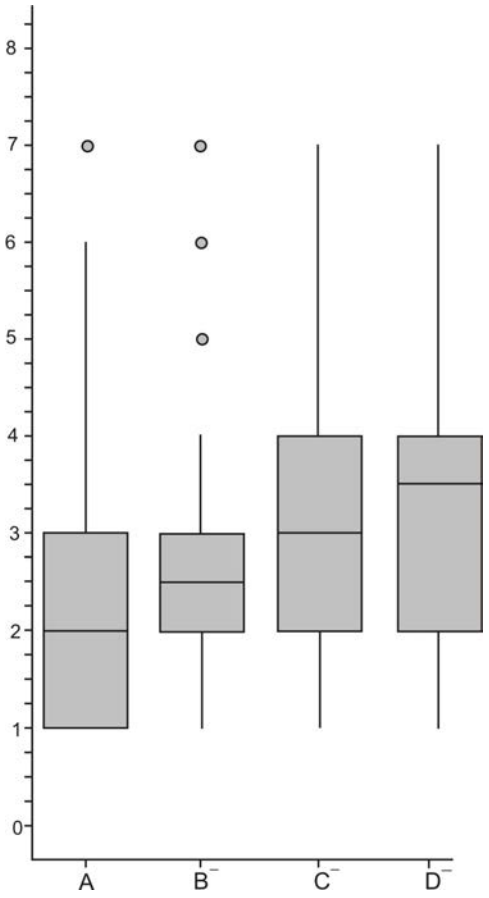
c) Maya asked her friends what their favorite movie was this year and graphed the results.

d) Jeno bought a large box of doughnut holes at the local pastry shop, weighed each of them, and then plotted their weights to the nearest tenth of a gram.

Questions 8 - 17 are multiple-choice questions. Select the best answer from the choices given.

8) Which of the following box plots matches the histogram?





9) Identify the 5 number summary for this set of numbers:

12,356; 16,564; 15,684; 12,358; 15,987; 13,556; 18,564; 18,965; 19,683; 18,432; 18,563; 19,352

a) {12,356; 14,600; 17,498; 18,000; 19,683}

b) {12,356; 14,620; 17,498; 18,764.5; 19683}

c) {12,356; 14,650.5; 17,498; 18,700.5; 19683}

d) {12,356; 14,683; 17,500; 18,800; 19683}

e) {12,356; 14,695.5; 17,900; 18,888; 19683}

10) Thirty students took a statistics examination having a maximum of 50 points. The grade distribution is given in the following stem-and-leaf plot:

0	9
1	225
2	013335889
3	00136679
4	02244478
5	0

The median grade is equal to:

a) 30.5

b) 30.0

c) 25.0

d) 28.5

e) 44.0

11) Ms. Davis conducted a survey of the 44 students in her stats classes and asked how tall each student is in inches. Here is the five-number summary of the students' data:

$\{57, 64, 67, 69, 79\}$

Approximately how many people are shorter than 64 inches tall?

a) 8

b) 21

c) 22

d) 11

e) 18

12) In which scenario(s) would it be better to use the 5-number summary versus the mean and standard deviation?

a) a graph that is skewed

b) a graph that is fairly symmetric

c) a graph that is symmetric but has several high outliers

d) Both choice (b) and (c)

e) Both choice (a) and (c)

f) All of (a), (b) and (c)

13) Suppose the lowest score on an English exam was 35% and the highest score was 90%. If the teacher of the class was to examine her students' test scores, which type of distribution would she prefer to see? One that is...

a) skewed to the right

b) skewed to the left

c) fairly symmetric

d) none of the above

14) Several people were surveyed as they were leaving a movie theatre. Among other things, they were asked how much money they had spent. They answers were: \$14, \$17.50, \$16, \$16, \$19.25, \$12.75, \$16, \$37.75, \$13.50 and \$17. It was later discovered that the person who answered "\$37.75" actually spent \$17.75. Which of the following would **not** change as a result?

a) the box plot

b) the mean & the mode

c) the median & the mode

d) the standard deviation

e) they all change

15) What does the following five-number summary tell you about the shape of the distribution? $\{5, 7.7, 9, 10.9, 24\}$

a) skewed to the right

b) skewed to the left

c) symmetric

d) uniform

e) cannot determine

16) According to the 1.5*(IQR) Criterion, what are the two cut-off values for determining whether the data set in question #15 contains any outliers?

a) 5 & 24

b) 7.7 & 10.9

c) 11.3 & 29.9

d) 4.5 & 14.1

e) 2.9 & 15.7

17) A class survey was conducted to determine students' preferences. One question regarded favorite sport to watch on TV. The results are as follows: 9 said "football"; 12 said "hockey"; 5 said "basketball"; 6 said "baseball" and 3 said "other". What would the central angle be for "hockey" in a pie chart of this data?

a) 65°

b) 123°

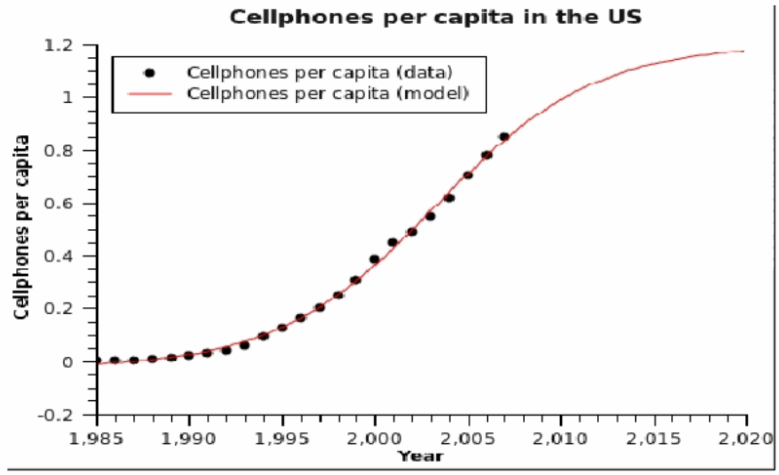
c) 90°

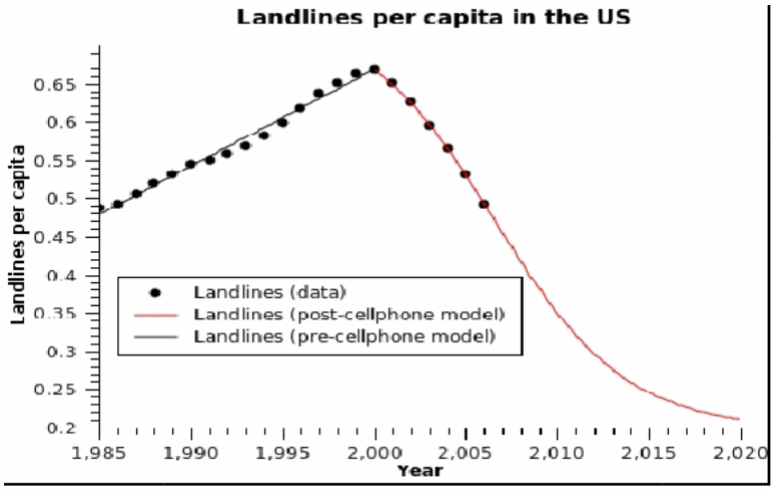
d) 34°

e) 111°

18) The following two graphs are based from the US Census Bureau, 2008 ('per capita' means per person). The dots represent actual data values, and the red curves represent models that can be used to predict future trends. Study the two graphs and answer the questions that follow.

Based on the number of cell phones in the US (US Census Bureau, 2008) and accounting for population the number of cell phones per capita, cp, can be modeled by the following graph:





Source:http://www.rationalfuturist.com August 2, 2011.

a) What type of graphs are these?

b) Describe the trend that each graph shows separately. This should be in context.

c) Notice that the horizontal scales are the same. Compare and contrast the trends that are shown in the two different graphs in context.

d) Approximately how many cell phones were there per person in 1997? In 2005? How many will there be, if the trend continues as the model indicates, in 2018?

e) Approximately how many landlines were there per person at the peak? What year did this occur? If the trend continues as the model indicates, how many landlines will there be per person in 2015?

19) The AHS Tornadoes and the BHS Bengals are big rivals! Every year students try to prove that their school is better at sports than the other school. The table below shows the number of points scored by each school's basketball team during the last 15 games played.

Table 5.12:

Tornadoes	Bengals
58	74
90	81
71	73
64	63
58	58
63	84
60	92
72	38
48	77
59	84
72	95
62	66
57	70
64	68
49	72

a) Construct a back-to-back stem plot for the data.

b) Calculate the five number summary, mean and standard deviation for both teams.

c) Construct parallel box plots for the data.

d) Compare the two distributions. This should be done in context and include at least three distinct comparisons.

e) What other information would you like to know when comparing these two basketball teams? Explain.

20) Here are the percents of the population that are of Hispanic origin in each state according to the 2010 Census.

State	Percentage	State	Percentage	State	Percentage
AL	3.9	LA	4.2	OH	3.1
AK	5.5	ME	1.3	OK	8.9
AZ	29.6	MD	8.2	OR	11.7
AR	6.4	MA	9.6	PA	5.7
CA	37.6	MI	4.4	RI	12.4
CO	20.7	MN	4.7	SC	5.1
CT	13.4	MS	2.7	SD	2.7
DE	8.2	MO	3.5	TN	4.6
FL	22.5	MT	2.9	TX	37.6
GA	8.8	NE	9.2	UT	13.0
HI	8.9	NV	26.5	VT	1.5
ID	11.2	NH	2.8	VA	7.9
IL	15.8	NJ	17.7	WA	11.2
IN	6.0	NM	46.3	WV	1.2
IA	5.0	NY	17.6	WI	5.9
KS	10.5	NC	8.4	WY	8.9
KY	3.1	ND	2.0	DC	9.1

Source:http://www.census.gov

a) Construct a histogram.

b) Calculate the five number summary.

c) Identify any outliers. Use the outlier test.

d) Accurately sketch a box plot.

e) What is the range? The IQR? The mode?

f) Calculate the mean and standard deviation.

g) Compare the mean and the median. *(i.e. which is larger? How different are they?)*

h) In this case would the 5#-summary or the mean & standard deviation be more appropriate? Why?

i) Describe the distribution. Be thorough! Don't forget your S.O.C.C.S! (shape, outliers, center, context, & spread)

j) According to the Census data, the percent of the total population of the United States that is Hispanic or Latino is 16.3%. Where does this number fit in the distribution of all of the states? Does it seem surprising? Why or why not?

21) An employer in Minneapolis was interested in determining how much money his employees were spending on parking each week. An SRS of 50 employees was selected to complete a questionnaire about parking. Several questions were asked including where they park, how much they spend per week, how often they have difficulties finding spots, if they pay daily, weekly, or monthly, etc. The following table is the average weekly expenditure for parking for this sample of 50 employees.

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	25
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

a) Construct a split-stem plot

b) Calculate the five number summary.

c) Identify any outliers. Use the outlier test

d) Accurately sketch a box plot. (to scale with labels)

e) What is the range? The IQR? The mode?

f) Calculate the mean and standard deviation.

g) Compare the mean and the median.

h) In this case would the 5#-summary or the mean & standard deviation be more appropriate? Why?

i) Describe the distribution. Be thorough! Remember your S.O.C.C.S!

Image References:

Gasoline.http://www.education.vic.gov.au

Cars.http://www.icoachmath.com

School Lunch pictogram.http://alex.state.al.us

Dot plot.http://cwx.prenhall.com

Stem plot example.http://www.basic-mathematics.com

Shapes of distributions.http://thesocietypages.org

Weight of Jessica graph. http://www.stat.psu.edu

Crowded stem plot. http://illuminations.nctm.org/

Three histogram example. http://classes.cec.wustl.edu

Package weight graph. http://flylib.com

Chapter 6

Analyzing Bivariate Data

6.1 Displaying Bivariate Data



Introduction

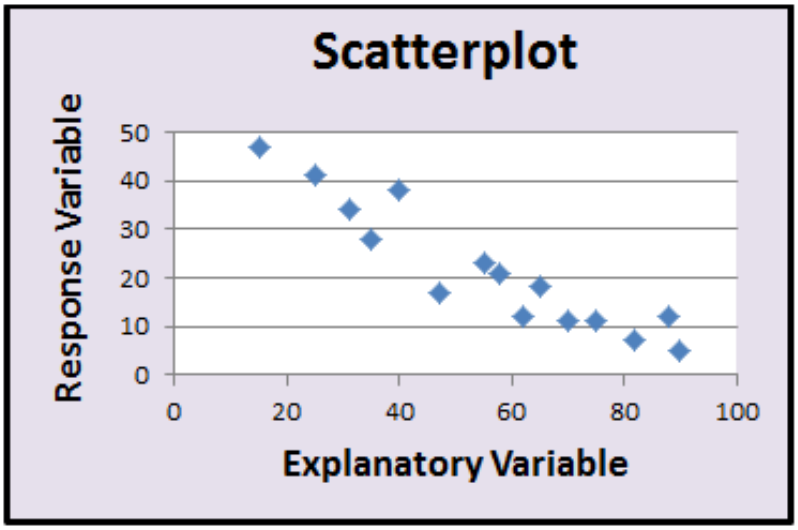
In chapter 5 we learned how to analyze and describe univariate, or single-variable data. We explored ways to present our data visually with graphs and charts and how to analyze our data with numerical statistics. Also, we described our findings verbally and in context. Now we will be analyzing bivariate numerical data. This means two numerical values collected about each individual. Such bivariate data is often given in a table, or can be listed as ordered pairs. We will construct appropriate graphs, calculate numerical statistics and equations, and describe the relationship between the two variables in context. The purpose will be to explore whether or not a relationship or association exists between the two numerical variables. If an association does exist, statistics can be used to predict one variable based on the other variable.

Learning Objectives

- Construct and interpret scatterplots
- Identify explanatory and response variables
- Describe bivariate distributions in context—including strength, outliers, form and direction

Scatterplots

Scatterplots are graphs that represent a relationship between two variables. Two numerical values are measured about each individual being studied. When these two values become ordered pairs that are graphed on a coordinate plane, the resulting graph is called a **scatterplot**. We often suspect that one of these variables might explain, cause changes in, or help to predict the other variable. The **explanatory variable** is the variable that we believe may explain or affect the other variable. The explanatory variable is plotted along the x-axis. The **response variable** is the variable we believe may respond to, or be affected by, the other variable. The response variable is plotted along the y-axis. The response variable is often referred to as the dependent variable and the explanatory is referred to as the independent variable. Even though we often look for an explanatory-response relationship between the two variables, we can create a scatterplot even if no such relationship exists.



Example 1

State whether or not you suspect that there will be an explanatory-response association between each of the following pairs of data. If yes, identify the explanatory and response variables.

a) A college professor decided to examine whether or not there is a relationship between the amount of time that a student studies and his or her score on the mid-term exam. At the end of the exam each student was asked to record the number of hours he or she had spent studying for the mid-term. The professor then made a scatterplot to examine the data.

b) A different professor wanted to see whether or not there is an association between her students' heights and their IQ scores. She gave each of her students an IQ test and had her TA measure each student's height to the nearest inch. She constructed a scatterplot to examine the data.

Solution

a) *It is reasonable to believe that the amount of studying does somehow have an effect on*

students' exam scores. The explanatory variable is hours studying and the response variable is exam score. Often thinking in terms of a cause and effect relationship can help identify which variable is which. As a hint, try to determine if one of the variables comes first. If one comes first, then is most likely the explanatory variable. In our example, studying should come before the exam.

b) *It is not reasonable to believe that there is an association between height and IQ scores. Neither of these variables comes before the other and neither would be useful in predicting the other. However, even though we do not believe that there is an explanatory-response relationship between these variables, we can still construct a scatterplot*

Example 2

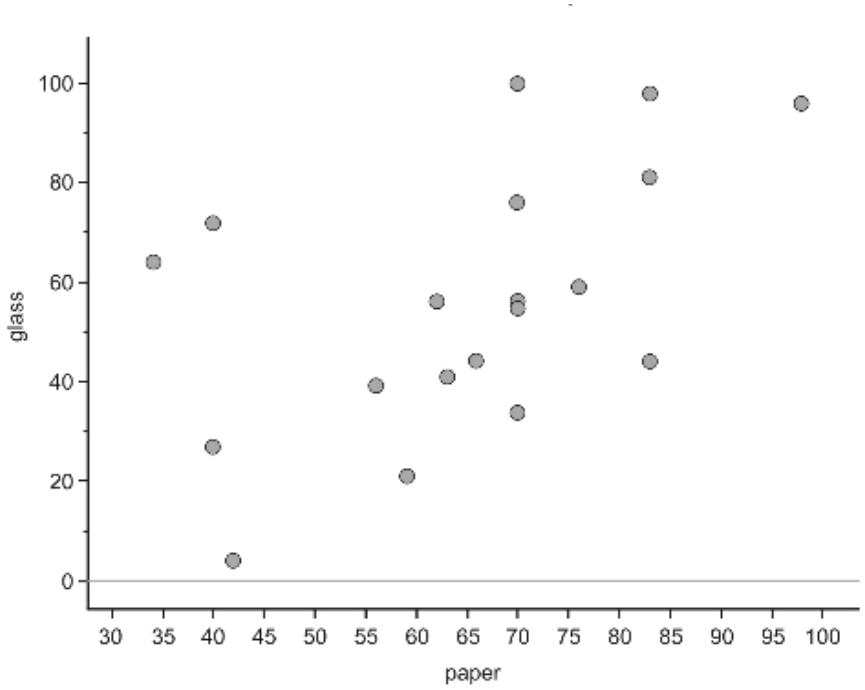
The following table reports the recycling rates for paper packaging and glass for several individual countries. It would be interesting to see if there is a predictable relationship between the percentages of each material that countries recycle. Construct a scatter plot to examine the relationship. Treat percentage of paper packaging recycled as the explanatory variable.

Country	%of Paper Packaging Recycled	%of Glass Packaging Recycled
Estonia	34	64
New Zealand	40	72
Poland	40	27
Cyprus	42	4
Portugal	56	39
United States	59	21
Italy	62	56
Spain	63	41
Australia	66	44
Greece	70	34
Finland	70	56
Ireland	70	55
Netherlands	70	76
Sweden	70	100
France	76	59
Germany	83	81
Austria	83	44
Belgium	83	98
Japan	98	96

Figure: Paper and Glass Packaging Recycling Rates for 19 countries

Solution

We will place the paper recycling rates on the horizontal axis because we are treating it as the explanatory variable. Glass recycling rates are then plotted along the vertical axis. Next, plot a point that shows each country's rate of recycling for the two materials. Be sure to label your axes.



Notice that we do not always need to start at zero on either axis when making scatterplots.

Describing Bivariate Data

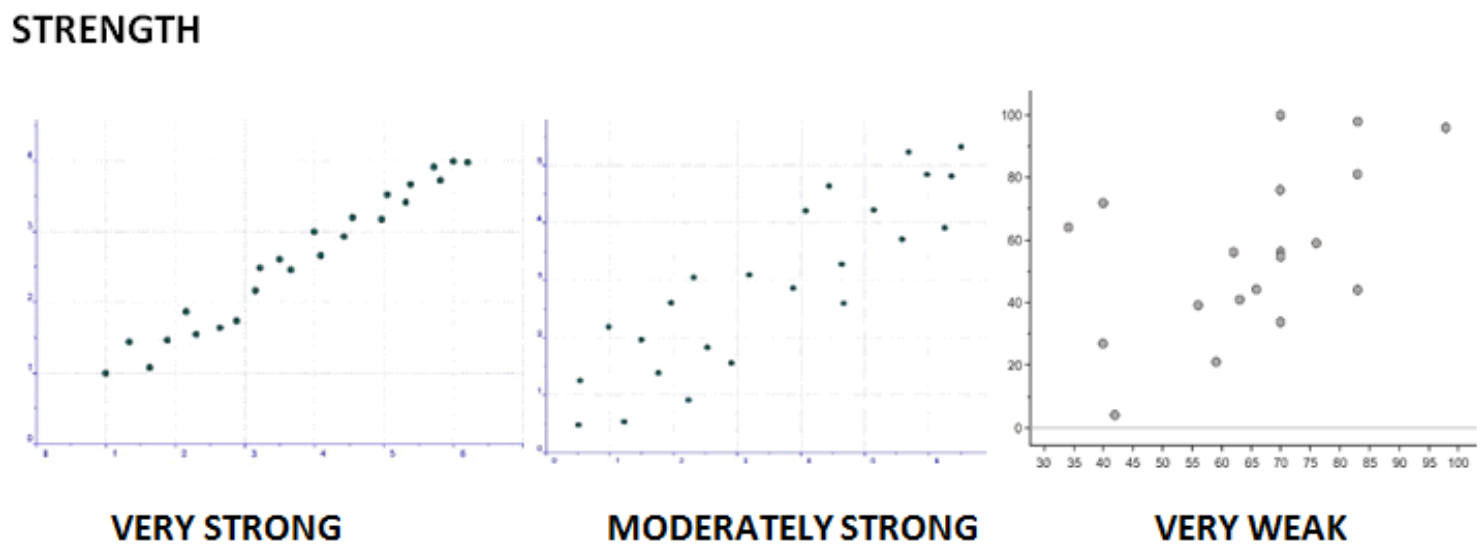
When we describe single variable data, we address several characteristics. We used the acronym *S.O.C.C.S.* to help remember to describe the shape, outliers, center and spread of a distribution. And, to be sure to do all of this in the context of the variables and individuals being studied. For bivariate data, we will again be discussing several characteristics in context. The important characteristics to describe when looking at the relationship between two numerical variables will be strength, outliers, form and direction. And, we will do this in the context of the variables and individuals being compared. The acronym that will help us to remember what to include in our descriptions is: **S.C.O.F.D.** (strength, context, outliers, form and direction).

When looking at a scatterplot, it is helpful to imagine drawing a line-of-best-fit through the data. A **line-of-best-fit** is a line that follows the trend of the data. It may go through some, all, or none of the actual points on the scatterplot. Do not actually draw such a line on your plot- just try to determine whether or not such a line would make sense, and if so, where it would fit. As you observe a scatterplot and imagine drawing such a line, you can ask yourself questions such as: *How close to a line do the points lie? Would a curved pattern fit better? Are there points that would be far away from the line? Would the line have a positive or negative slope? etc.*

Strength

Once you have constructed a scatterplot, you can examine the strength of the relationship between the two variables. The **strength** refers to how closely the points form a pattern. The more closely the points fit a pattern, the stronger the relationship between the variables. The more spread out and scattered the points are, the weaker the relationship. The first plot shows an extremely strong, linear pattern because the points form an obvious line. The second plot is more scattered so it is only moderately strong. And,

the third plot does not show much of a pattern at all, so it is very weak. Keep in mind that the association may be very strong, but not linear. We could find a very clear curved pattern in the data, for example. In the next section of this book we will learn about a statistic, called correlation, that measures the strength of the linear relationship between two variables.



In example #4, the relationship between paper and glass recycling rates for these countries is very weak.

Context

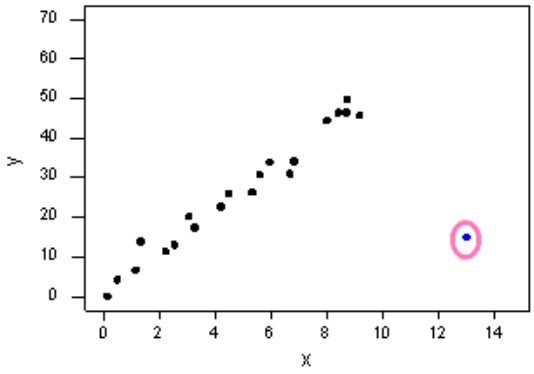
Do not forget that the graph, the numbers and equations, and the descriptions are all about something-its **context**. All of these elements should be described in the context of the variables and the individuals being examined.These graphs and statistics are not meaningless, they are about something!

In example #4, the scatterplot explores the relationship between glass and paper recycling rates for several countries..

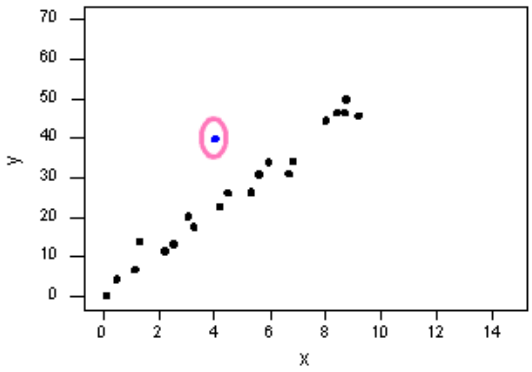
Outliers

When examining a scatterplot, look for any data values that do not fit the pattern, or points that stand out from the rest of the data. An **outlier** will be a point that lies away from the rest of the data or one that seems to affect the strength of the relationship between the two variables. Many outliers will weaken the association between the variables, but they often would not significantly change where a line-of-best-fit would be drawn. An **influential point** is an outlier that actually seems to influence the line-of-best-fit. Imagine what the plot would look like without the point in question. If it would change the strength, then the point is an outlier. If it would change the slope of a line-of-best-fit, or where the line would be drawn, then the point is influential.

OUTLIERS



OUTLIER & INFLUENTIAL



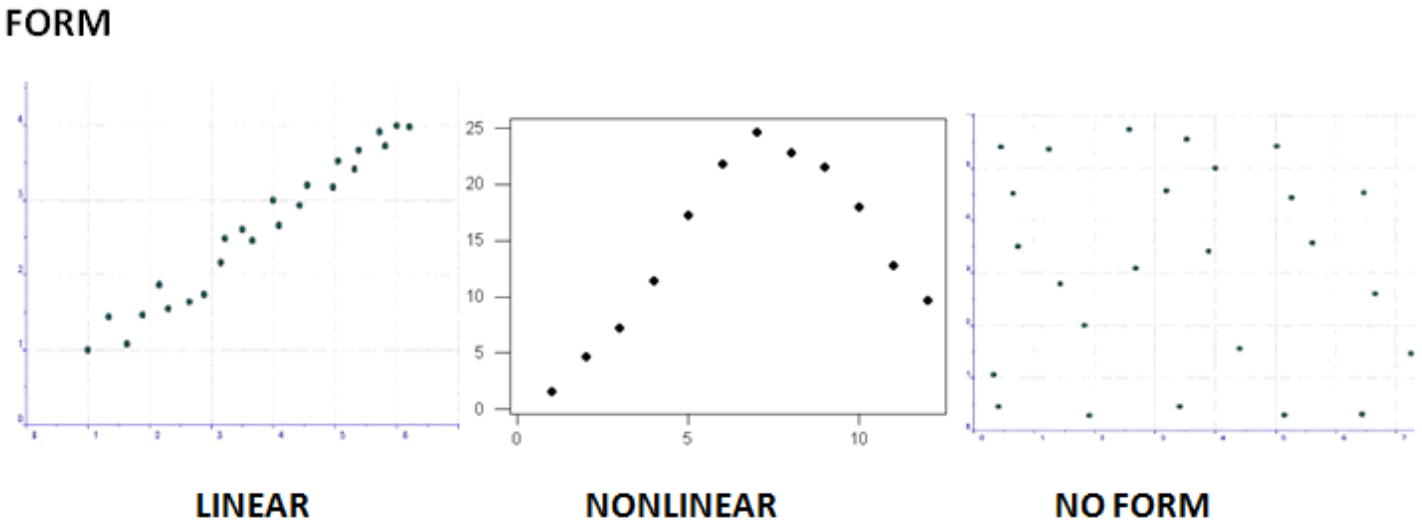
OUTLIER (but not influential)

In example #4, there seem to be some outliers. For example, Estonia and New Zealand have much lower paper recycling rates than their glass rates. Without these data values, the relationship would be stronger.

Form



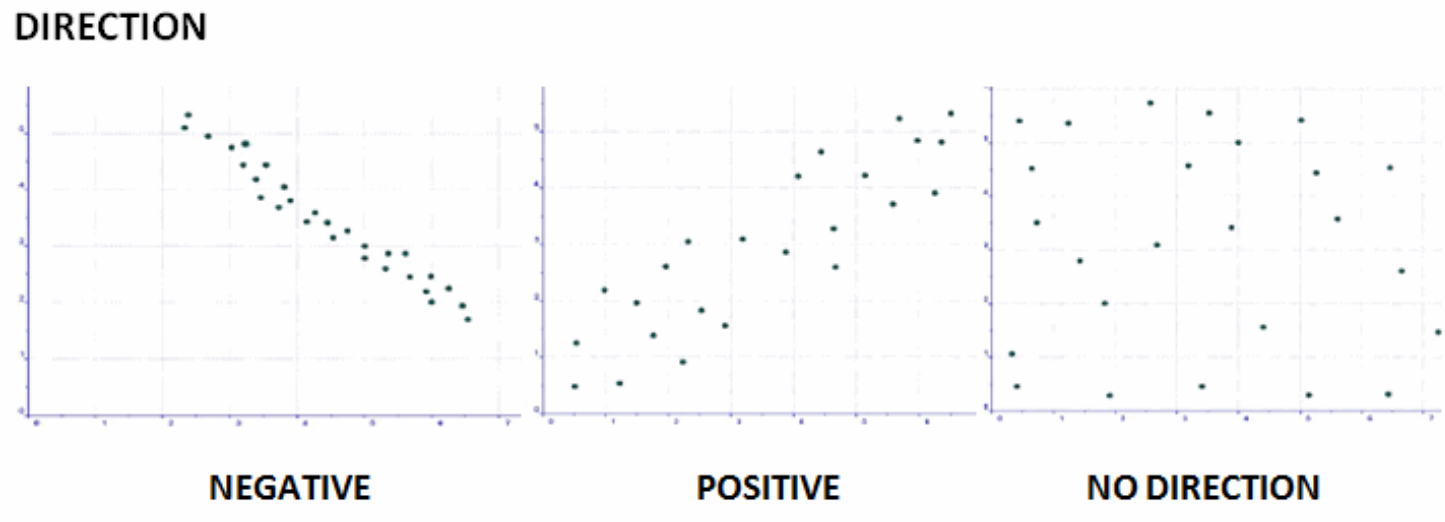
Many scatterplots show a clear **form** or pattern. The first plot below shows a clearly linear. It is easy to imagine drawing a line-of-best-fit through these points. The second plot shows a clearly curved form. A line would not make any sense, so this is non-linear. The third plot shows a great deal of scatter among the points, so it has no form whatsoever.



In example #4, the scatterplot for paper and glass recycling rates shows a very weak linear form. The relationship is very weak, but no curved pattern is visible. If the outliers were removed, it would become more linear.

Direction

The direction of the graph is also important to mention. A graph that goes down to the right has a **negative association.** That is, as the explanatory variable increases, the response variable decreases. The first plot below has a negative relationship between the variables. A graph that goes up to the right has a **positive association**. That is, as the explanatory variable increases, the response variable also increases. The second plot shows a positive relationship between the variables. The third plot is an example of a graph that has neither a positive, nor a negative direction. If the relationship is linear and a line-of-best-fit is added to the graph, the slope of the line will be positive if the association is positive. And, the line will have a negative slope if there is a negative linear association between the two variables.



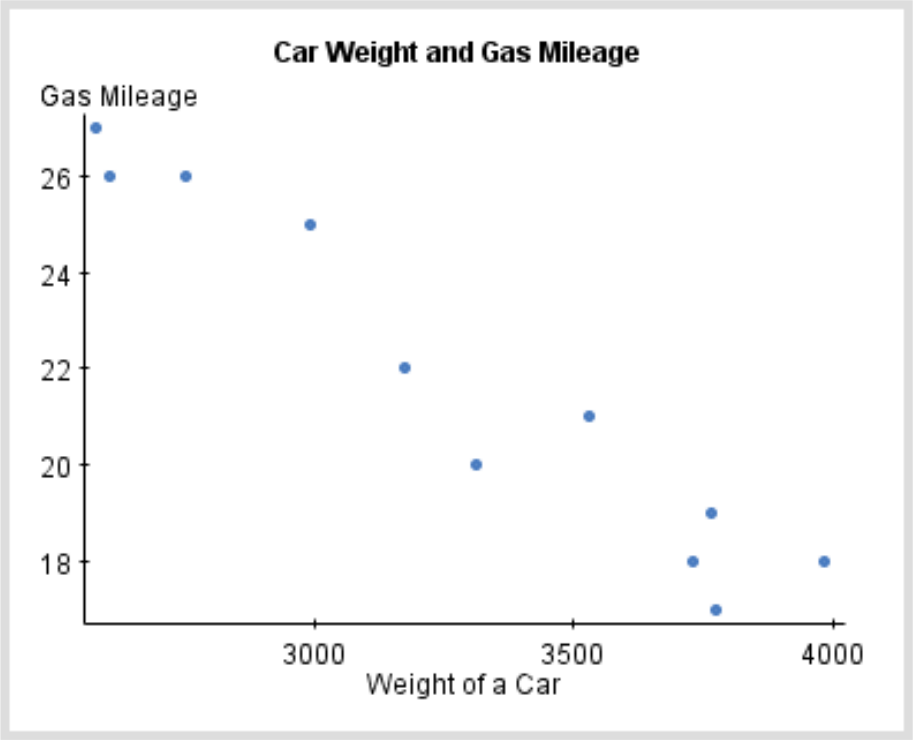
In example #4, the scatterplot for paper and glass recycling rates shows a positive association. As the paper recycling rate for these countries increases, so does the glass recycling rate.

Example 3

The following example is a scatterplot showing the weights and gas mileage for several cars.

a) Identify the explanatory and response variables.

b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).



Solution

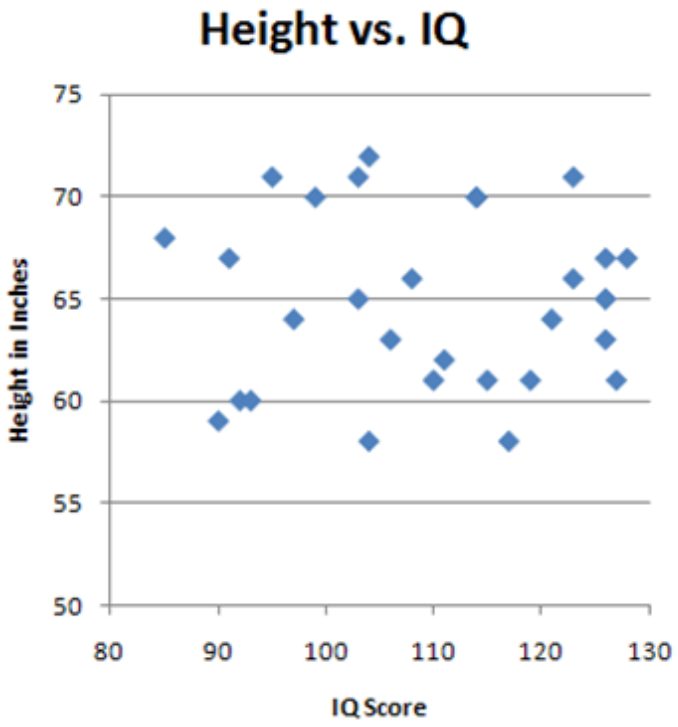
a) explanatory variable is: ***weight of the cars***

response variable is: ***gas mileage of the cars***

b) ***The relationship between these vehicles' weights and gas mileage is strong and very linear. There are no extreme outliers visible in the graph. The association between a vehicle's weight and gas mileage is negative. As the weight of the vehicles increase, the gas mileage of the vehicles decrease.***

Example 4

The following scatterplot shows the data collected by the professor who wanted to see whether or not there is an association between her students' heights and their IQ scores. She gave each of her students an IQ test and had her TA measure each student's height to the nearest inch. Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).



Solution

There appears to be no relationship between height and IQ scores for these students. The graph has no form an no direction. Therefore, there are no outliers. The relationship has zero strength. There is no pattern or trend between IQ scores and students' heights.

Section 6.1 Exercises

1) State whether or not you suspect that there will be an explanatory-response association between each of the following pairs of data. If yes, identify the explanatory and response variables.

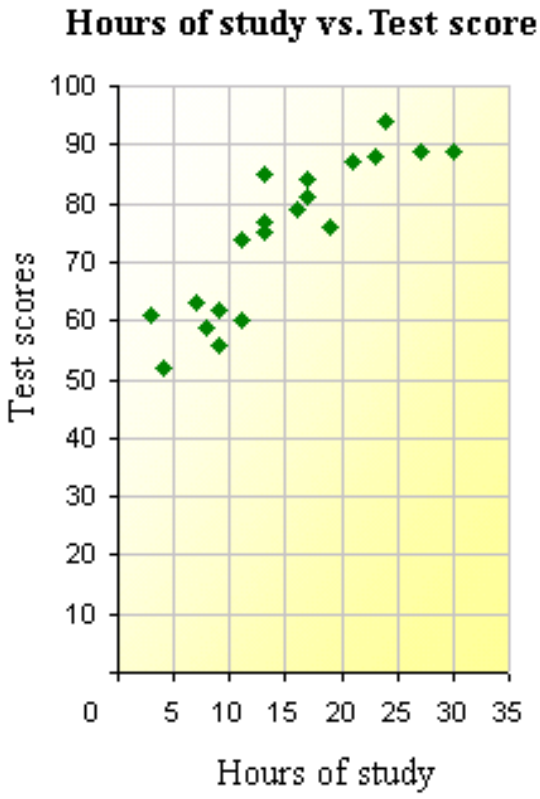
a) The number of semesters that students have been enrolled in college and the number of credits that they have earned.

b) Students' grades on a statistics test and their weights.

c) Employees' annual salary and the number of years that they have been employed by the company.

d) The number of songs each person has on his or her IPod and the number of months that they have owned the IPod.

2) A college professor decided to examine whether or not there is a relationship between the amount of time that a student studies and his or her score on the mid-term exam. At the end of the exam each student was asked to record the number of hours he or she had spent studying for the mid-term. The professor then made a scatterplot to examine the data. Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).



3) Malia turned the water on in her bathtub full blast. She then measured the depth of the water every two minutes until the bathtub was full (and her mother started to freak out). Her findings are listed in the following table.

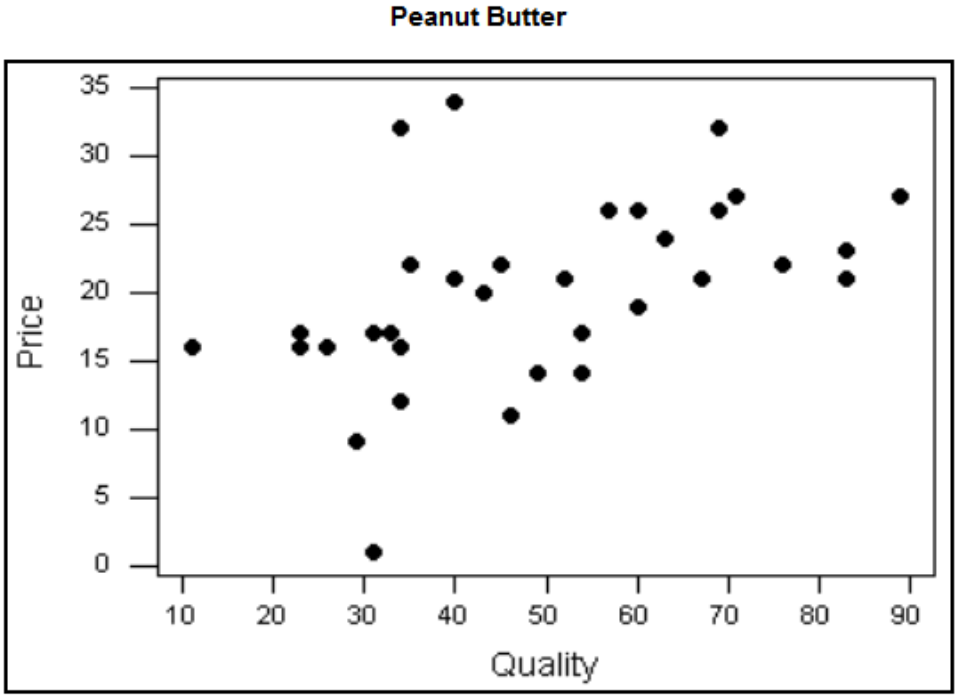
Time (minutes)	Depth (cm)
2	7
4	9.5
6	14
8	19.5
10	21
12	24
14	32
16	36
18	37.5
20	41
22	46

a) Identify the explanatory and response variables for this situation.

b) Construct a scatterplot to show the results.

c) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

4)Several brands of peanut butter were rated for quality. The following graph compares the price per ounce and the quality rating for each of these brands of peanut butter.



a) Identify the explanatory and response variables for this situation.

b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

5) Mr. Exercise wanted to know whether or not customers continued to use their equipment after they purchased it. He contacted an SRS of his customers who had purchased an exercise machine during the past 18 months. His findings are summarized in the following table:

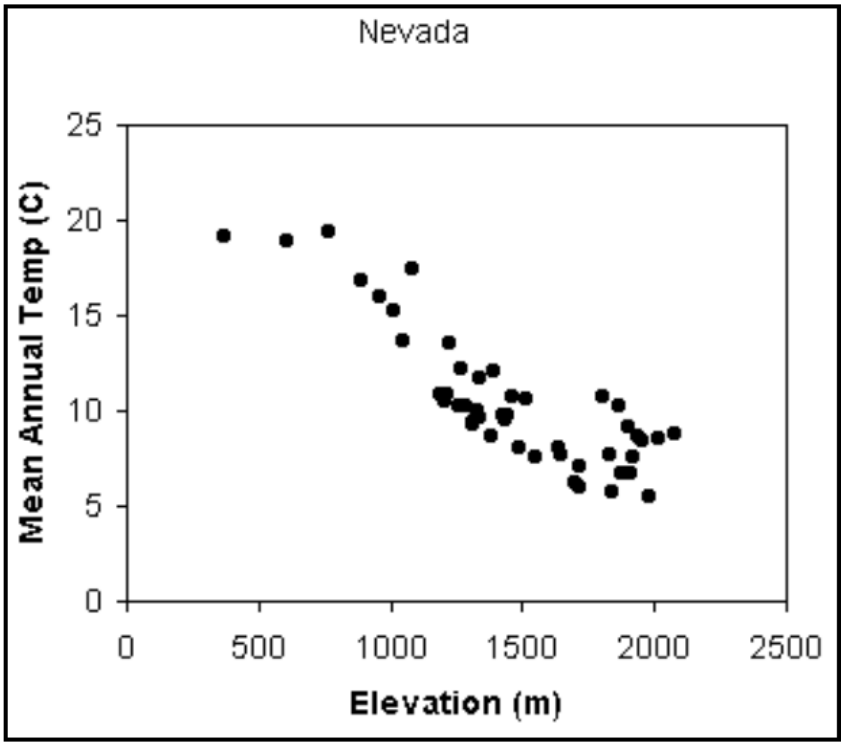
# months owned machine	# hours exercise per week
1	8
5	4.5
7	3
4	6
9	2
14	1.5
5	7
11	4
3	6.5
6	4

a) Identify the explanatory and response variables for this situation.

b) Construct a scatterplot to show the results.

c) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

6) The following scatterplot shows the elevation and mean temperature for various location in Nevada.



a) Identify the explanatory and response variables for this situation.

b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

6.2 Correlation

Learning Objectives

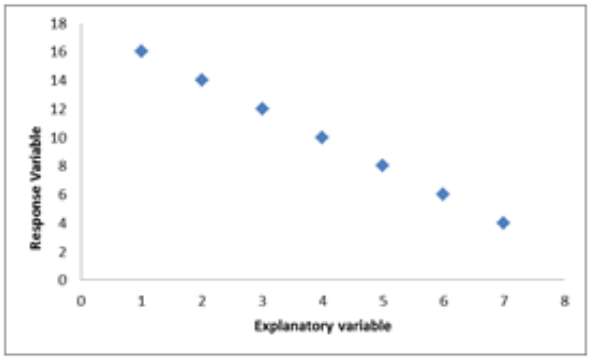
- Understand the properties of the linear correlation coefficient
- Estimate and interpret linear correlation coefficients
- Understand the difference between correlation and causation
- Identify possible lurking variables in bivariate data
- Understand the effects outliers and influential points can have on correlation

The Correlation Coefficient

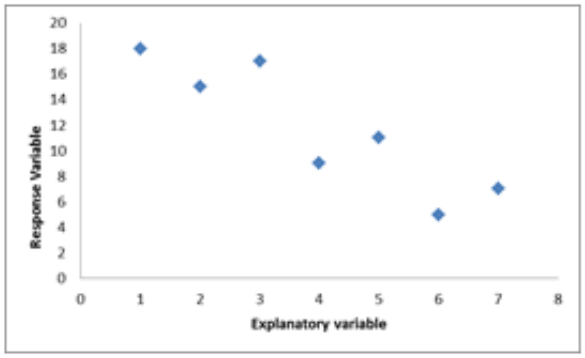
The **correlation coefficient** is a statistic that measures the strength and direction of a linear relationship between two numeric variables. The symbol for correlation is **r**, and r can take any value from -1.0 to +1.0. The correlation coefficient (r) tells us two things about the linear relationship between the two variables, its strength and its direction. The **direction** of the relationship, positive or negative, is given by the sign of the r value. A positive value for r indicates that the relationship is positive (increasing to the right), and a negative r value indicates a negative relationship between the two variables (decreasing to the right). Bivariate data with a positive correlation tells us that as the explanatory variable increases, so does the response variable. And, bivariate data with a negative correlation tells us that as the explanatory variable increases, the response variable decreases. A correlation of zero indicates neither of these trends.

The second thing that the correlation coefficient tells us is the **strength** of the linear relationship - how close the points are to forming a perfect line. An r of exactly 1 or -1 has a perfect correlation, the relationship forms a perfect, exact line. An r value of exactly +1 means that the relationship forms a perfect line with a positive slope and a r value of exactly -1 means that the scatterplot will show a perfect line with a negative slope. The closer the correlation value is to either +1 or -1, the stronger the linear relationship is. And, as r gets closer to zero (either positive or negative), the weaker the linear relationship is. It is important to note that this is only measuring the linear relationship between the two variables. If the relationship shows a clear curved pattern for example, the correlation will tell us nothing about the strength of the relationship.

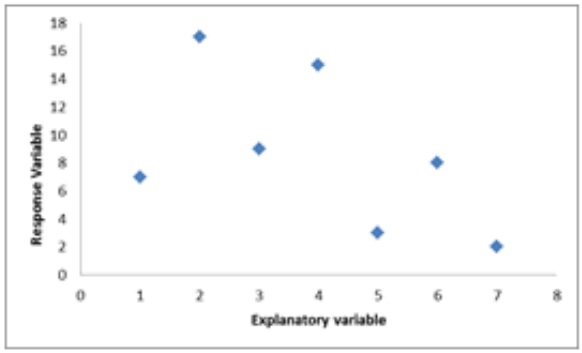
Here are some sample scatterplots with their correlation coefficients given:



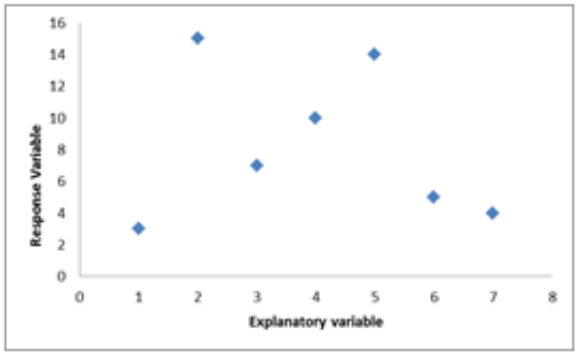
r = -1



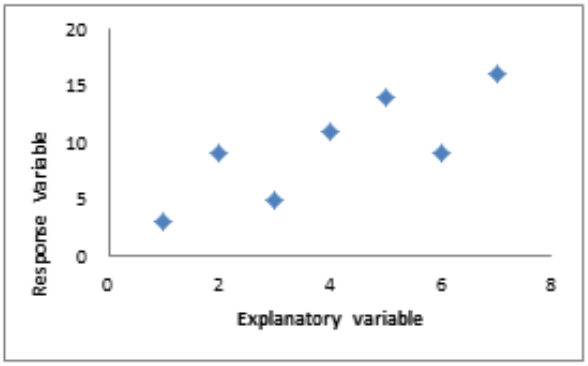
r = -.900



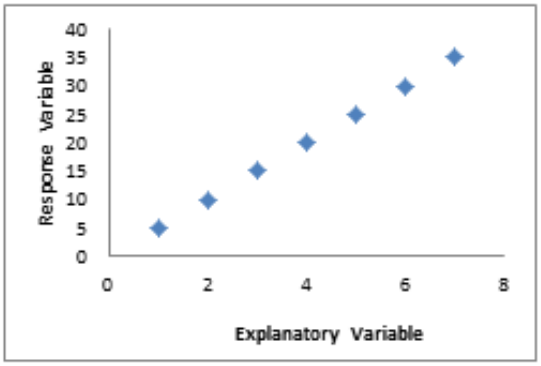
r = -0.536



r = -.160



r = +0.803



r = +1

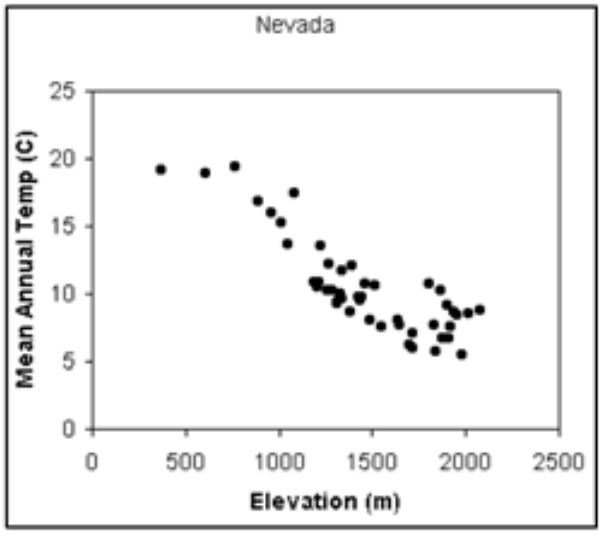
We will be using either our calculator or a computer to **calculate** the correlation coefficient. The formula to calculate the correlation coefficient is quite tedious. It involves calculating the mean and standard deviation of all of the x-values and the mean and standard deviation of all of the y-values. It then compares the x-value of each ordered pair to the mean of x and every y-value to the mean of y (by subtracting and then dividing by the standard deviation), multiplies these newly calculated values, adds all of them, and divides by one less than the sample size. The correlation formula is shown below, but we will be using technology rather than calculating by hand.

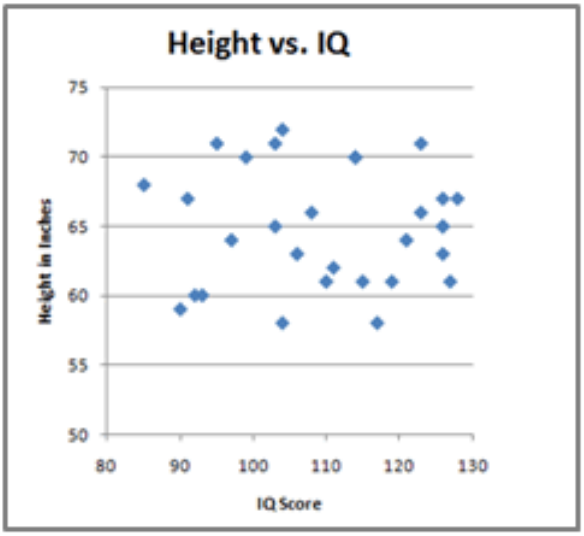
Correlation Coefficient Formula:

$$r = \frac{\sum \left(\frac{(x_i - \bar{x})}{s_x}\right)\left(\frac{(y_i - \bar{y})}{s_y}\right)}{n - 1}$$

Example 1

Estimate the correlation coefficient for each of the following scatterplots.





Solution

Nevada: The correlation will be negative and fairly strong, so my estimate is r ≈ -0.85.

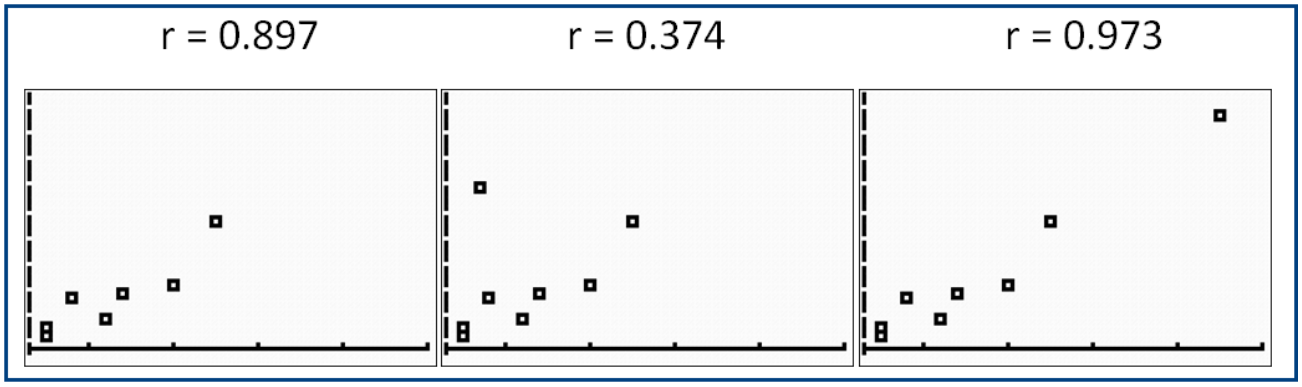
Height & IQ: There seems to be no pattern to the graph, so my estimate is r ≈ 0.

Properties of Correlation

When considering using correlation as a measure of the strength between two variables, you should construct and examine a scatterplot first. It is important to check for outliers, be sure that the relationship appears to be linear, be sure that your sample size is sufficient, and consider whether the individuals being examined

were too much alike in some way to begin with. Thus, when examining correlation, there are four things that could affect our results: outliers, linearity, size of the sample and homogeneity of the group.

An outlier, or a data point that lies outside of our overall pattern, can have a great effect on correlation. How great of an affect is determined by the sample size of the data and by the magnitude by which the outlier lies outside of the pattern. The three plots below show scatterplots with their correlation coefficients (r). The first plot shows a positive and reasonably linear graph. Its correlation is r = .897, which is positive and fairly strong. The second plot shows the same data as plot one, with one outlier (upper left) added. Its correlation has dropped to r = .374, which is still positive, but much weaker. This demonstrates how outliers can bring the correlation closer to zero. However, some outliers can actually strengthen the correlation. This is demonstrated in the third plot, which shows the same data as the first with one outlier (upper right) added. With this outlier, the linear relationship becomes even stronger than the first plot, at r = .973.



If the relationship is not linear, calculating the correlation coefficient is meaningless. It is only testing the linear relationship between the two variables. Imagine a scatterplot that shows a perfect parabolic relationship. We would know that there is a strong relationship between these two variables, but if we calculated the correlation coefficient, we would arrive at a figure around zero. Therefore, the correlation coefficient is not always the best statistic to use to understand the relationship between variables.

As we discussed in experimental design, a small sample size can be misleading. It can either appear to have a stronger or weaker relationship than is really accurate. The larger the sample, the more accurate of a predictor the correlation coefficient will be on the linear relationship between the two variables.

When a group is too much alike in regard to some characteristics (homogeneous), the range of scores on either or both variables is restricted. For example, suppose we are interested in finding out the correlation between IQ and salary. If only members of the Mensa Club (a club for people with IQ's over 140) are sampled, we will most likely find a very low correlation between IQ and salary since most members will have a consistently high IQ, but their salaries will vary. This does not mean that there is not a relationship – it simply means that the restriction of the sample limited the magnitude of the correlation coefficient.



Lurking Variables

It is very important to know that a high correlation does not mean causation! Often times studies that showing a high correlation between two variables will influence readers into thinking that one variable is the cause of the relationship. This is not always true! A high correlation simply does not prove that one

variable is causing the other. In some situations we would agree that one variable is in fact causing the response in another. The best way to prove such a direct **cause-and-effect** relationship is by carrying out a well designed experiment. For example, smoking is strongly correlated with lung disease, and, based on much scientific evidence, we can now say that cigarette smoking causes lung disease. However, this topic was highly debated for many years before the surgeon general announced that it was accepted that cigarette smoking causes lung cancer and emphysema. Many people refused to accept this for many years. People who stood to lose money if smoking was proven to be unsafe, suggested every possible other explanation that they could think of. They suggested that it was simply a coincidence, or that all people who choose to smoke might have something else in common that was actually the cause of the lung disease, not the cigarettes. Because it was not ethical to experiment on humans in order to prove the direct cause-and-effect relationship, the debates went on for a long time.

Congress mandated that the Surgeon General's warning labels appear on all cigarette packaging sold in the U.S. beginning in January 1966. Since 1972, the Surgeon General's warning labels have appeared on U.S. cigarette advertising as well. The Surgeon General's warnings required by the Cigarette Labeling and Advertising Act of 1965 have been amended over time, and the Act currently requires cigarette manufacturers and importers to print the following warnings, which are rotated on a quarterly basis, on cigarette packaging and advertisements:


SURGEON GENERAL'S WARNING: Smoking Causes Lung Cancer, Heart Disease, Emphysema, And May Complicate Pregnancy.
SURGEON GENERAL'S WARNING: Quitting Smoking Now Greatly Reduces Serious Risks to Your Health.
SURGEON GENERAL'S WARNING: Smoking By Pregnant Women May Result in Fetal Injury, Premature Birth, And Low Birth Weight.
SURGEON GENERAL'S WARNING: Cigarette Smoke Contains Carbon Monoxide.

Sometimes the relationship between variables is a cause-and-effect one, but many times it can be simply a coincidence that the two variables are highly correlated. It is also possible that some other outside factor, a **lurking variable**, is causing both variables to change. A situation where we have two variables that are both being affected by some other, outside, lurking variable is called **common response**. For example, we can show a high correlation between the number of TV's per household and the life expectancy per person among many countries. However, it makes no sense that TV's cause people to live longer. Some lurking variable is having an effect here. It is likely that the economic status of the countries is causing both variables to change: more money means more TV's and more money means better health care. If a country is wealthy it is much more likely to have citizens who own TV"s. Also, if a country is wealthy it is much more likely to have good hospitals, roads, health education, access to clean water and food, all things that contribute to longer life.

In some situations we will have two variables that are highly correlated, but we are unsure of the exact cause of the relationship. We may be unclear as to whether or not one is causing the other, if there is a lurking variable causing a common response, or if there is some unknown lurking variable that is related in some other unknown way (lurking variables are not always obvious to the researchers). Such a situation is called **confounding**, because it is confusing to determine how the variables are related (if at all), and whether there may be some lurking variable and if it is related to the variables in question. The variables seem all mixed up and the relationship is unclear, even if highly correlated. An example of confounding is global warming. This is a highly debated topic in social media and web-blogs. Some people argue that human pollution is a major cause of the increase in CO_2 and other green house gasses in the atmosphere. While others argue that it is a part of a natural cycle that has normally occurred in our Earth's history. Still some may think both explanations are at work. This is an example of confounding because there is confusion about the cause of global warming.

And don't forget that some relationships are occurring completely by chance, and their high correlation is then just a **coincidence**. For example, if you researched divorce rates and gas prices over the past 50 years you may note that both have gone up. A scatterplot comparing divorce rates and gas prices would show a strong positive relationship. The correlation would likely be a high, positive value. However, it makes no sense that divorce rates are causing high gas prices. It also is unlikely that there exists a common response or some form of confounding. So in this case, we would say that this is a relationship that is best explained by sheer coincidence.

Example 2

Suggest possible lurking variables to explain the high correlations between the following variables. Explain your reasoning. *Consider whether common response, confounding, or coincidence may be involved.*

a) It has been shown that cities with more police officers also have higher numbers of violent crimes. Does this mean that more police officers are causing more violent crimes to occur?

b) Over the past 25 years, the percent of parents using car-seats has increased significantly. During this same time period, the rate of DUI arrests has also increased significantly. These two variables, when graphed, show a very high, positive correlation. Does this mean that car-seat use is causing DUI's to increase?

c) A study published in *USA Today* claimed that, *"Teens who text a lot [are} more likely to try sex, drugs, alcohol."* Does this mean that texting causes teens to try sex, drugs and alcohol? Could we then limit teen behaviors such as these by canceling their texting plans?

Solutions

a) It makes no sense that the number of police officers would be causing the violent crime to occur. It is much more likely that it the reverse, that communities with high numbers of violent crimes need higher numbers of police officers. It is also probable that both variables increase in cities with higher populations. Due to the fact that we can think of more than one possible lurking variable and it is difficult to know how all of these variable actually relate, we would say that this is an example of confounding (the variables in question and the lurking variables are all mixed up).

b) It is clearly ridiculous to think that car-seat use is causing an increase in the rate of DUI's. It also makes no sense that DUI's cause car-seats to be used. It may be simply a coincidence that these are both increasing. Or, perhaps there has been in increase in law enforcement for both over this time period. The awareness of the dangers of both have increased over the past 25 years, so maybe this is an example of common response. Or, maybe many factors contribute to the increase of both, so perhaps this is an example of confounding. But, no matter what, this is not cause-and-effect.

c) It is unlikely that texting is actually the cause of these behaviors. There is most likely some other, lurking variable(s) that are the cause(s). One probable lurking variable, when it comes to teenagers, is the parents. Perhaps this is an example of a common response to parents who are not very involved in their teens' lives. Parents who are not very involved would not be aware that their teen is texting too much and would also not be aware of what choices their teen is making during his or her free time. Perhaps teens who spend a lot of time unsupervised would be more likely to text and would also be more likely to try sex, drugs, and alcohol. All of

these behaviors might be a common response to not having parents who prohibit or limit teens from doing these things. Canceling texting plans would have little to no affect on other teen behaviors.

See the link for more information on this report at: **http://www.usatoday.com/yourlife/sex-relationships/2010-11-10-texting-teens_N.htm**

Multimedia Links:

Calculating Correlation on the Internet,

There are several websites where you can enter in data points and find their correlation one of them is below.

http://easycalculation.com/statistics/correlation.php

If this site no longer works, trying googling "finding correlation applet" and see what you get for results.

For an explanation of the correlation coefficient,

see kbower50, The Correlation Coefficient (3:59).

Another, more lighthearted example of Correlation ≠ Causation can be found at the following website, which discusses the evil of the pickle.

http://www.exrx.net/ExInfo/Pickles.html

For a better understanding of correlation try these fun links below,

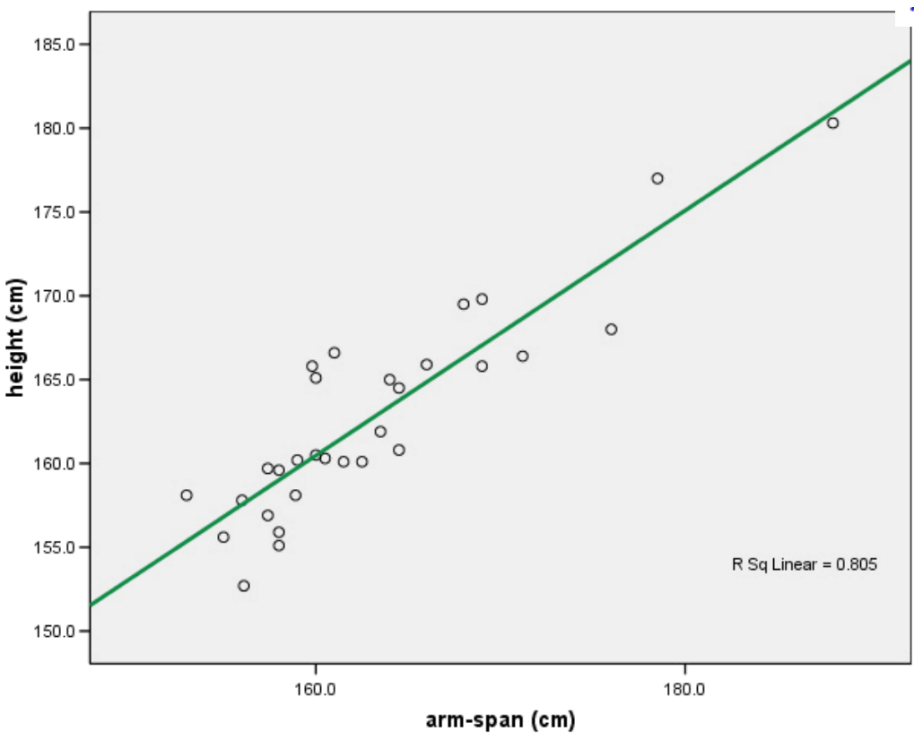
http://www.istics.net/stat/Correlations Match the graph to its correlation.

http://www.rossmanchance.com/applets/guesscorrelation/GuessCorrelation.html Guess the correlation Guess the correlation

Section 6.2 Exercises

1) What are the two things that the correlation coefficient measures?

2) The program used to create this scatterplot found the line-of-best-fit and reported the r-squared value as $r^2 = 0.805$ for the relationship between arm-span and height for several individuals. What is the correlation coefficient? Is it positive or negative? Explain how you know.



3) During the summer Ms. Statsteacher lets her two daughters stay up later than during the school year. Their bedtimes during the summer range from 8:30 p.m. to 12:30 a.m. She has discovered that her older daughter Reily will wake up between 8:00 and 9:00 a.m. no matter what time she goes to bed. However, her younger daughter Neila tends to wake up later after she gets to stay up later, and earlier when she goes to bed earlier. Neila has been known to wake up anytime between 8:00 and 11:45 a.m.

a) Sketch a separate (approximate) scatterplot for **each** daughter, that compares time going to sleep and time waking up. *Which will be explanatory and which will be response?*

b) Which of these do you think will best approximate the correlation for Reily?

A. close to r = +1
B. close to r = +.75
C. close to r = 0
D. close to r = -.75
E. close to r = -1

c) Which of these do you think will best approximate the correlation for Neila?

A. close to r = +1

B. close to r = +.75

C. close to r = 0

D. close to r = -.75

E. close to r = -1

4) Suggest possible lurking variables to explain the high correlations between the following variables. Explain your reasoning. *Consider whether common response, confounding, or coincidence may be involved.*

a) *As ice cream sales increase, the rate of drowning deaths increases sharply. Does this mean that ice cream causes drowning?*

b) With a decrease in the number of pirates, there has been an increase in global warming over the same time period. Does this mean global warming is caused by a lack of pirates?

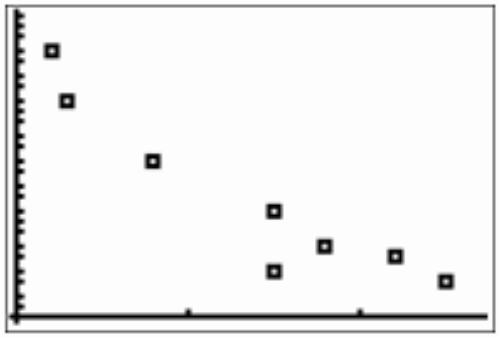
c) The higher the number of fire-fighters fighting a fire, the more damage done by the fire. Does this mean that we can limit damage by sending fewer fire-fighters to fires?

d) Suppose that each of the hockey players on the high school team supplies his or her own hockey stick, with varying degrees of flex. The assistant coach has been keeping a record of the degree of flex for each player's stick and their respective point totals (goals and assists). He has noted that there is a strong, negative correlation between these two variables. In other words, the players with less flex in their sticks are scoring more points and those with more flex are scoring fewer points. Can we then give players less flexible sticks and expect to increase scoring?

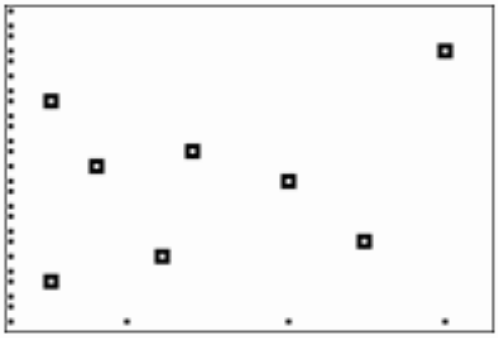
5) In a recent study in *Resource Manual*, it was noted that divorced men were twice as likely to abuse alcohol as married men. The authors concluded that getting divorced caused alcohol abuse. Do you agree? Explain your reasoning.

6) A commercial for a new diet pill claims *"You will lose weight while you sleep! No exercise needed!"*. They then show several before-and-after photos of people who have lost weight. People who were obese are now very buff. They then give the information for you to order the pills *("for three payments of just $19.95 each, plus shipping and handling")*. Is this proof that these diet pills caused these people to lose weight? Suggest possible lurking variables. Explain your reasoning.

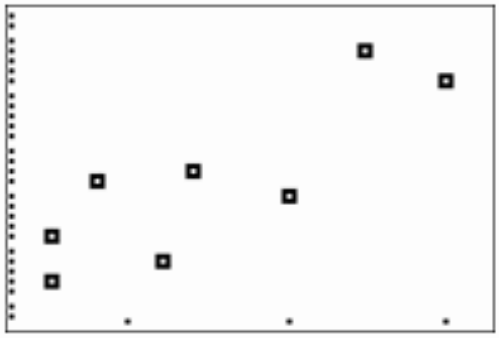
7) Match each graph with its correlation coefficient:



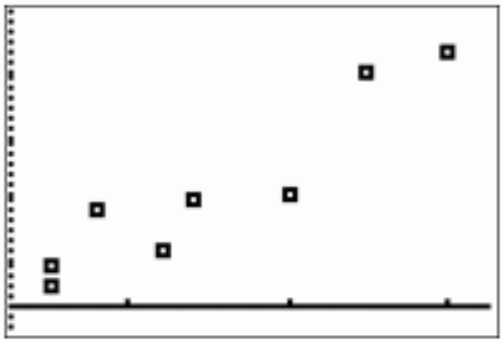
GRAPH #1



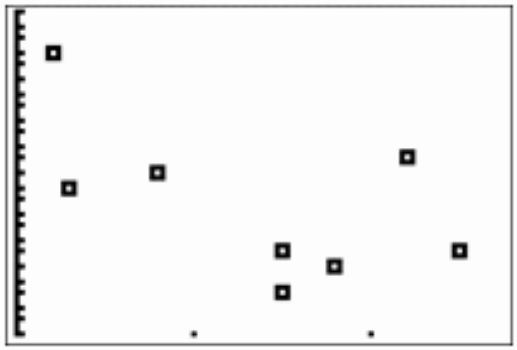
GRAPH #2



GRAPH #3



GRAPH #4



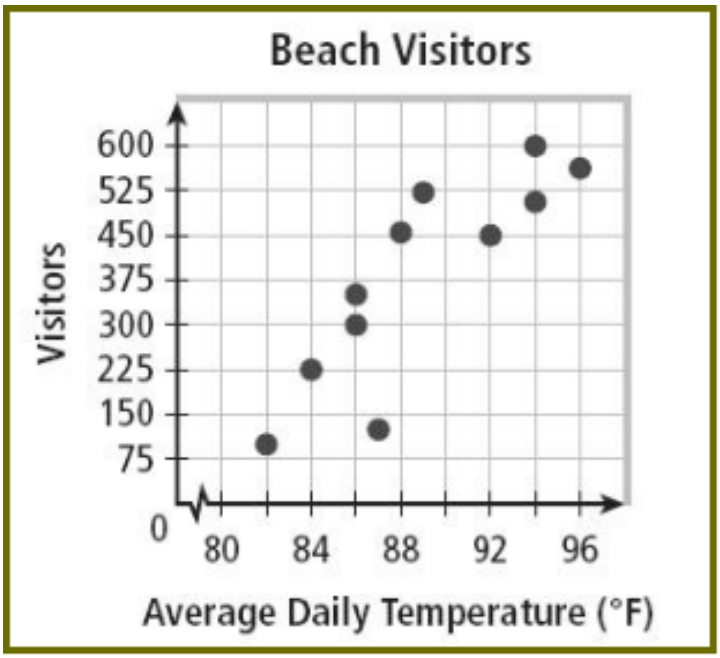
GRAPH #5

Match the Correlation with the graph:

A. r = 0.941
B. r = 0.850
C. r = 0.321
D. r = -0.598
E. r = -0.938

8) A correlation of r = 0 indicates no linear relationship between the two given variables. But, this does not mean that there is no relationship between the two variables. Sketch a scatterplot in which there is a strong relationship between the variables, but the correlation would be near r = 0.

9) Use the "Beach Visitors" scatterplot to answer the questions that follow.



a) Identify the explanatory and response variables.

b) Estimate the correlation coefficient for the graph.

c) Describe what the scatterplot shows. (*remember S.C.O.F.D*)

6.3 Least-Squares Regression



Learning Objectives

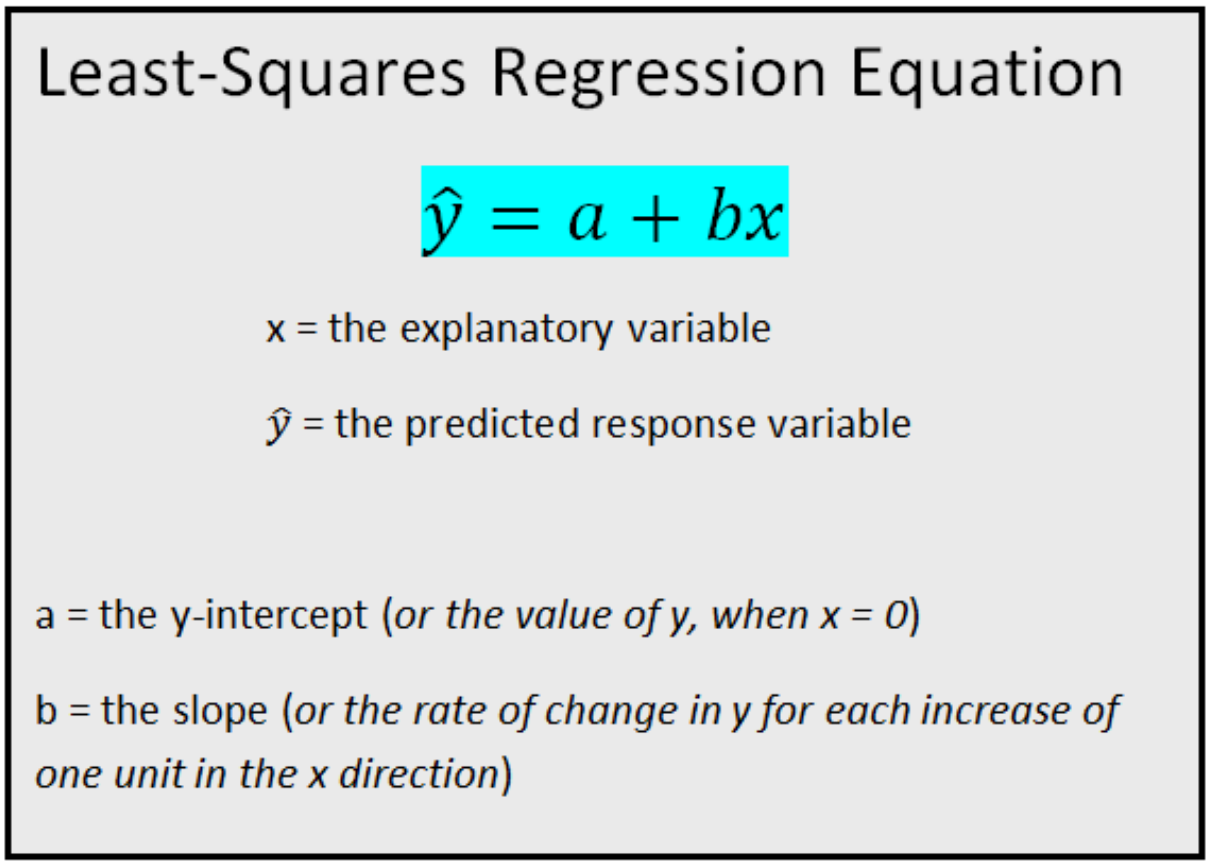
- Construct scatterplots using technology
- Calculate and graph the least-squares regression line using technology
- Calculate the correlation coefficient using technology
- Use the LSRL to make predictions
- Understand interpolation and extrapolation
- Interpret the slope and the y-intercept of the LSRL

Least-Squares Regression

In the last section we learned about the concept of correlation, which we defined as the measure of the linear relationship between two numerical variables. We saw that when the points of a scatterplot formed a clear linear pattern, then the points were said to have a high correlation. As a reminder, we can have a strong correlation in either a positive (increasing to the right) or a negative (decreasing to the right) direction. We have also discussed the idea of drawing a line-of-best-fit through the data. In some scatterplots this is easy to do and all of us would end up with our lines in nearly the same place. However, if everyone were to simply draw a line where they think it fits or to select two of the points to calculate a line through, our lines and equations would certainly vary from person to person. Therefore, we will use a specific formula to calculate the equation for the line-of-best-fit.

Linear regression involves using data to calculate a line that best fits the data and then using that line to predict scores. We will use the **Least-Squares Regression Line (LSRL)-** the line that makes the sum of the squares of the vertical distance of each data point from the line the least possible. This is the standard regression equation that is used most often. It is the one that your graphing calculator and Excel will calculate for you. The formula and process to calculate this is quite tedious, so we will use technology to find the LSRL equations. The regression equation will be in the form of: $\hat{y} = a + bx$, where **a** is the y-intercept and **b** is the slope of the equation. Your calculator will calculate the correlation coefficient (**r**) at the same time as it calculates the LSRL equation. It will also report a value for r^2 (which is exactly what it says; r-squared). The $\mathbf{r^2}$ value is called the coefficient of determination, but we will not be addressing its importance in this course.

To calculate the LSRL equation and correlation coefficient, use a graphing calculator or computer program. See the appendix at the end of this book for the steps to calculate the LSRL and correlation.


Least-Squares Regression Equation
$\hat{y} = a + bx$
x = the explanatory variable
$\hat{y}$ = the predicted response variable
a = the y-intercept (*or the value of y, when x = 0*)
b = the slope (*or the rate of change in y for each increase of one unit in the x direction*)

Interpreting the slope and y-intercept

As with all of our statistics, these data, graphs and equations are not meaningless. They represent the relationship between two numerical values measured on several specific individuals. Thus the slope and the y-intercept of our newly calculated regression equation mean something as well. So, we will be interpreting both in context. The **interpretation of the slope** of the regression equation is the rate of change in the response variable (y), for each increase of one unit of the explanatory variable (x). You will say something like: *For each increase of one (explanatory variable), there will be (an increase or decrease) of (slope value) in the (response variable).*

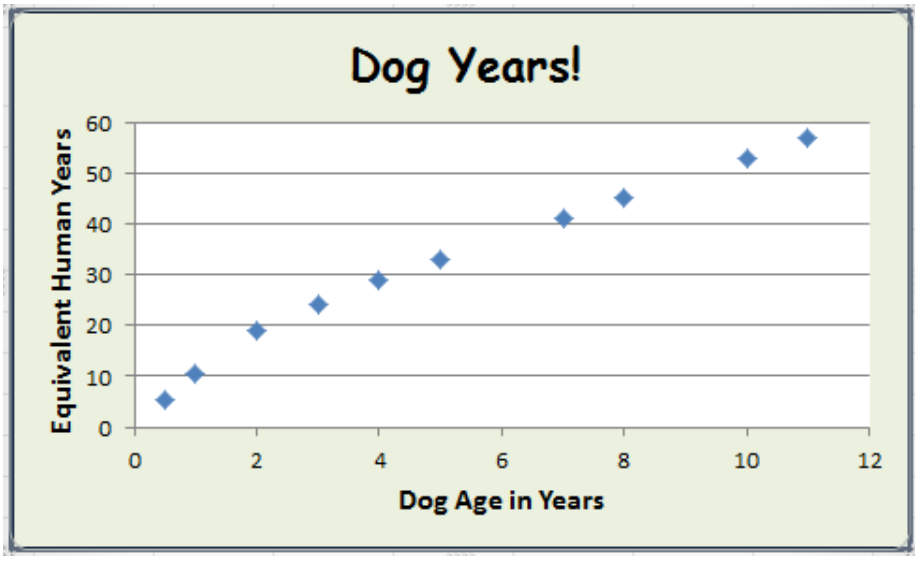
The **interpretation of the y-intercept** of the regression equation is the value of the response variable (y) when the explanatory variable (x) is zero. You will say something like: *When (explanatory variable) is zero, the (response variable) is (y-intercept value).* You will discover that the interpretation of the y-intercept often makes absolutely no sense when put into context. This is because actual data rarely involves x-values of zero.

Example 1

Below is data given by a canine expert. It relates a dog's age in years to what they believe the equivalent age in human years to be.

Dog Age (in Years)	Equivalent Human Age (in Years)
0.5	5.5
1	10.5
2	19
3	24
4	29
5	33
7	41
8	45
10	53
11	57

The scatterplot showing this data, using dog age as the explanatory variable, is below.



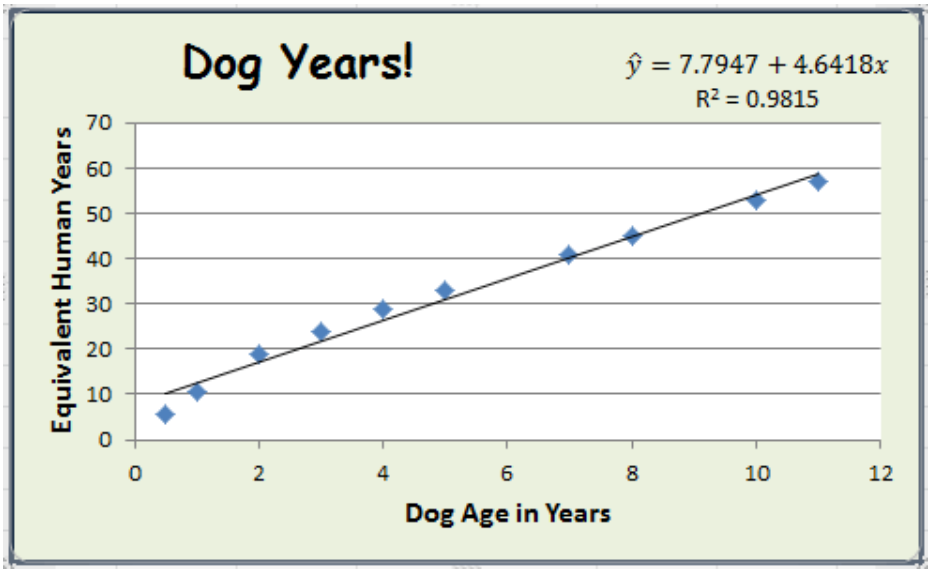
a) Calculate the Least-Squares regression line for the Dog Year Data. Report you equation. Be sure to identify your variables.

b) Calculate the correlation (r). What two things does r tell us about this relationship?

c) Identify and interpret the slope in the context of the problem.

d) Identify and interpret the y-intercept in the context of the problem.

Solution



*a)*This was done using Excel, but the graphing calculators will report the same LSRL.

LSRL is: $\hat{y} = 7.795 + 4.642x$

$x =$ *Dog age in years*

$y =$ *equivalent human years (predicted)*

b) r will be the square-root of r^2 (The graphing calculators report both r and r^2 so you would not need to do any calculating, but Excel only gave r^2).

$r = \sqrt{r^2} = \sqrt{0.9815} = 0.9907$

The two things that r tells us are: ***Because r is positive, this relationship is increasing. And r is very close to one, so this relationship is very strong.***

c) ***The slope is 4.642. It means that for every increase of one year in dog age, there is an increase of 4.642 years in the equivalent human age.***

d) ***The y-intercept is 7.795. It means that if a dog were 0 zeros old, it would be 7.795 years in human years. (This is clearly nonsense in this case!)***



Making predictions

The main use of the regression line is to predict values. After calculating this line, we are able to predict values by simply substituting a value for the explanatory variable (x) and solving the equation for the predicted response value (y). In our example above, we can predict that the human year equivalence for a dog that is 6 years old is approximately 35.6 human years (see equation below). This prediction is reasonable and it matches with our graph. However this is not always the case.

$$\hat{y} = 7.795 + 4.642(6) = 35.647$$

As you look at the LSRL drawn on the above scatterplot, you can see that the points to the far left do not appear to be very linear. So, using the line to the left of about 1 year will not make much sense. Also, we do not have any idea what will happen to the data beyond the 11 years that we have recorded. An LSRL is very useful in making predictions, but only within the range of the actual data that we have collected and can see- this is called **interpolation**. We can see that this line is a reasonably good fit between 1 and 11 dog years, but we simply do not know what happens beyond 11 years (and we cannot use negative years for obvious reasons). The prediction line that we have calculated will go forever in both directions (remember geometry?), but it will not be appropriate to use it to predict for all values of x. Using a regression line to predict values that are outside the range of our actual data is called **extrapolation**. Extrapolation will often yield ridiculous answers! However, even if the result seems reasonable, we should avoid extrapolating because we simply do not know what happens beyond our actual observations. Making decisions based on extrapolating can be dangerous as we are coming to conclusions that are not backed up by data.

Example 2

The following table lists the GPA and Verbal SAT Score for seven students. Analyze how well Verbal SAT Scores can be used to predict students' GPAs based on this data.

Student	Verbal SAT Score	GPA
Anna	595	3.4
Bryce	520	3.2
Corbin	715	3.9
Delia	405	2.3
Emilio	680	3.9
Frankie	490	2.5
Geraldine	565	3.5

a) Construct a scatterplot on your graphing calculator (or computer). Sketch the graph that the calculator shows. Be sure to label your axes.

b) Calculate the Least-Squares Regression Line (LSRL) using your calculator. Report your equation. Be sure to identify your variables.

c) Calculate the correlation coefficient (r). Report it here. What are the two things that this number tells us about this graph?

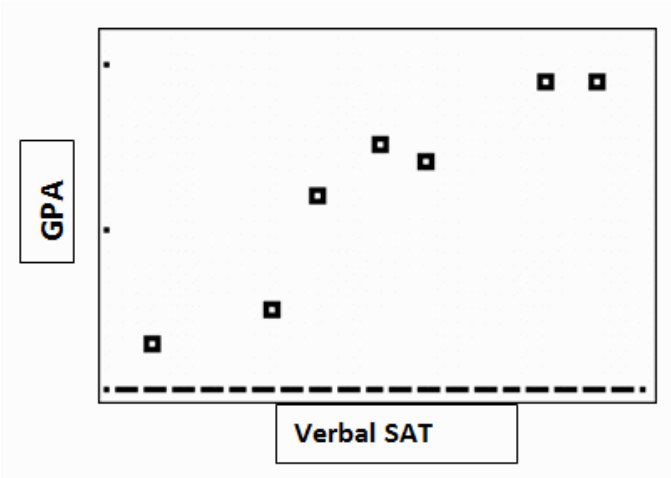
d) Identify and interpret the slope in the context of the problem.

e) Using your equation, what is the predicted GPA of a student who has a Verbal SAT Score of 500? Of a student with a score of 600?

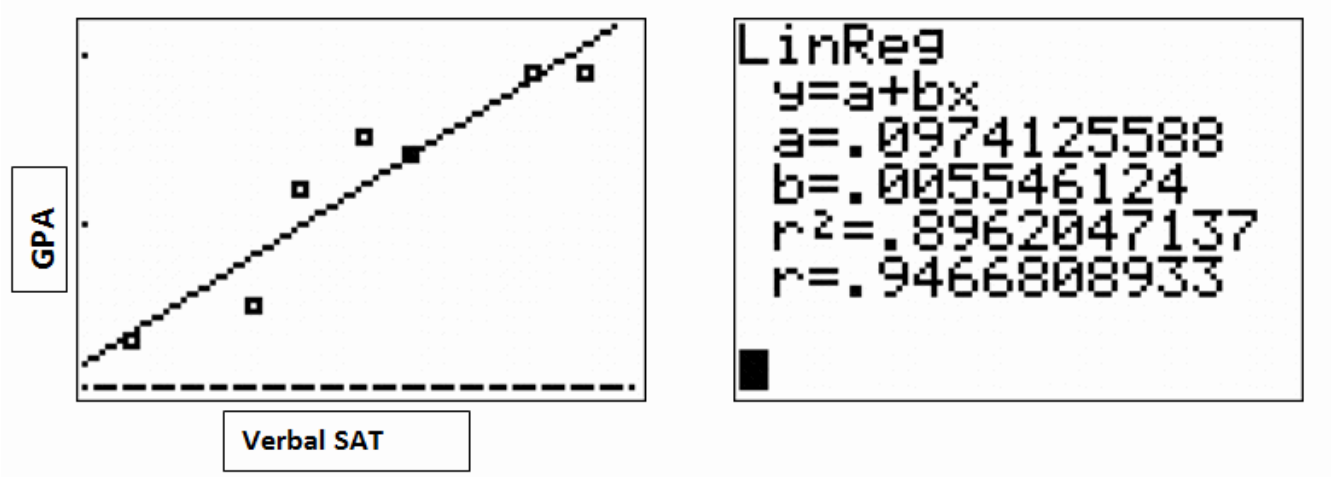
Solution

a) Construct a scatterplot on your graphing calculator (or computer). Sketch the graph that the calculator shows. Be sure to label your axes.

Here is the scatterplot from a TI-84 plus:



Here are the LSRL, correlation, and the scatterplot with the line added to the graph, from a TI-84 plus:



b) Calculate the Least-Squares Regression Line (LSRL) using your calculator. Report your equation. Be sure to identify your variables.

LSRL is: $\hat{y} = 0.097 + 0.0055x$

x = Verbal SAT Score

y = predicted GPA

c) Calculate the correlation coefficient (r). Report it here. What are the two things that this number tells us about this graph?

The correlation is r = +0.9467. This tells us that the relationship is positive and strong.

d) Identify and interpret the slope in the context of this problem.

The slope is 0.0055. This tells us that for each increase of 1 point on the Verbal SAT Score, there will be an increase of 0.0055 in a student's GPA.

e) Using your equation, what is the predicted GPA of a student who has a Verbal SAT Score of 500? Of a student with a score of 600?

$\hat{y} = 0.097 + 0.0055(500) = 2.847$

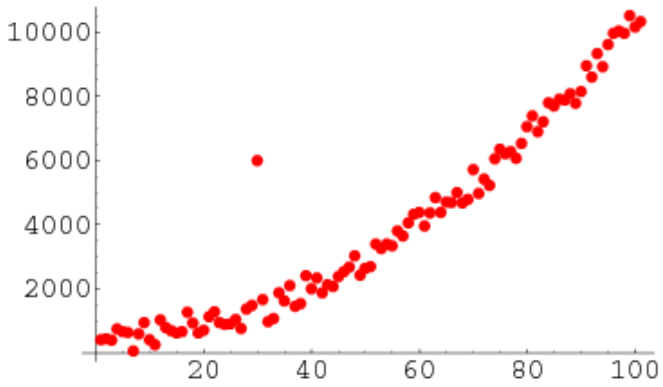
$\hat{y} = 0.097 + 0.0055(600) = 3.397$

So, the predicted GPA for a student who scores 500 on the SAT Verbal, is approximately 2.8.

And, the predicted GPA for a student who scores 600 on the SAT Verbal, is approximately 3.4.

Outliers and Influential points

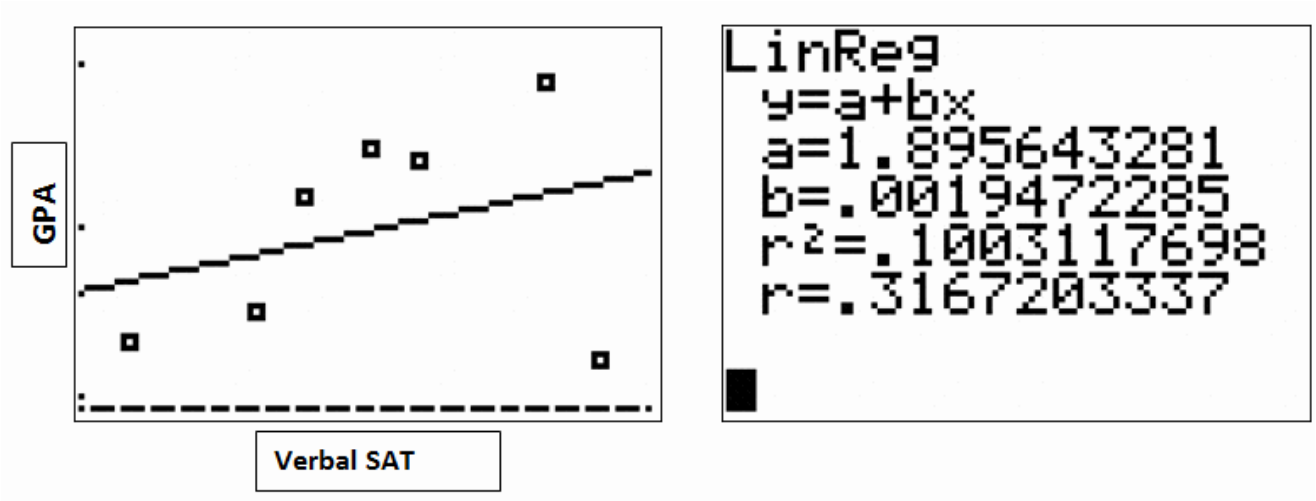
An outlier is an extreme observation that does not fit the general pattern of the data (see the example below). Because an outlier is an extreme observation, the inclusion of it may affect the correlation, and the equation for the least-squares regression line. When examining a scatterplot and calculating the regression equation, it is worth considering whether extreme observations should be included or not.



Let's use our GPA example to illustrate the effect of a single outlier. Suppose that we have a student who has scored very high on the SAT Verbal exam, but has a lower GPA. We will change Corbin's results to be 715 on the SAT and a GPA = 2.2, and see what happens to the LSRL and correlation.

Student	Verbal SAT Score	GPA
Anna	595	3.4
Bryce	520	3.2
Corbin	715	2.2
Delia	405	2.3
Emilio	680	3.9
Frankie	490	2.5
Geraldine	565	3.5

Here are the LSRL equation and the correlation coefficient recalculated with Corbin's GPA changed:



As you can see, this one change turned Corbin into an outlier. This caused the correlation to drop from r = 0.947, all the way down to r = 0.317. This is a huge change- it makes the relationship between the two variables extremely weak (rather that very strong). Also, this changed both the slope and the y-intercept of the LSRL equation dramatically. This means that predictions based on this LSRL will have different results than those based on the LSRL with Corbin's old GPA.

There is no set rule when trying to decide how to deal with outliers in regression analysis, but you can now see how an outlier really can change everything when it comes to scatterplots, correlation and least-squares regression. Be sure to mention any potential outliers that you observe in any scatterplot.

Section 6.3 Exercises

1) Malia turned the water on in her bathtub full blast. She then measured the depth of the water every two minutes until the bathtub was full. Her findings are listed in the following table. In an section 6.1 we constructed a scatterplot and described the plot, we are now going to analyze this data further.

Time (minutes)	Depth (cm)
2	7
4	9.5
6	14
8	19.5
10	21
12	24
14	32
16	36
18	37.5
20	41
22	46

a) Construct a scatterplot on your graphing calculator (or computer). Sketch the graph that the calculator shows. Be sure to label your axes.

b) Calculate the Least-Squares Regression Line (LSRL) using your calculator. Report your equation. Be sure to identify your variables.

c) Calculate the correlation coefficient (r). Report it here. What are the two things that this number tells us about this graph?

d) Identify and interpret the slope in the context of the problem.

e) Using your equation, what is the predicted depth of the water after 17 minutes? After one hour?

f) Are your answers in (e) reasonable? Why or why not?

2) The following table shows the progression of the Federal Minimum Wage in the United States since 1938 (source:http://www.laborlawcenter.com). We are going to analyze the relationship between year and minimum wage to see if there is a predictable relationship between the variables.

FEDERAL MINIMUM WAGE HISTORY

Effective Date	*Hourly Wage*
10/24/1938	$0.25
10/24/1939	$0.30
10/24/1945	$0.40
01/25/1950	$0.75
03/01/1956	$1.00
09/03/1961	$1.15
09/03/1963	$1.25
02/01/1967	$1.40
02/01/1968	$1.60
05/01/1974	$2.00
01/01/1975	$2.10
01/01/1976	$2.30
01/01/1978	$2.65
01/01/1979	$2.90
01/01/1980	$3.10
01/01/1981	$3.35
04/01/1990	$3.80
04/01/1991	$4.25
10/01/1996	$4.75
09/01/1997	$5.15
07/24/2007	$5.85
07/24/2008	$6.55
07/24/2009	$7.25

a) Using **year only** as the explanatory variable (ignore month & day), construct a scatterplot. Sketch the graph that the calculator shows. Be sure to label your axes.

b) Describe the relationship between the two variables. (S.C.O.F.D.)

c) Calculate the Least-Squares Regression Line (LSRL). Add the line to your graph and report your equation. Be sure to identify your variables.

d) Calculate the correlation (r). Even though r is very high, do you feel that a line is the best model for this data? Why or why not?

e) Based on your model, what would you predict the Federal Minimum Wage to be in 2012? Is this an accurate prediction? Why or why not?

f) Based on your model, what would you predict the minimum wage to have been in 1968? How close is this to the actual minimum wage that year?

3) Suppose that some researchers analyzed the relationship between fathers' and sons' IQ scores for a group of men. Suppose further that they discovered that the relationship was reasonably linear and they calculated a regression line of $\hat{y} = 12 + 0.9x$; where x= father's IQ and y = son's IQ.

a) Identify the explanatory and response variables.

b) Identify and interpret the slope in the context of the problem.

c) Identify and interpret the y-intercept in the context of the problem.

d) Do your answers to (b) and (c) seem reasonable? Why or why not?

e) What would you predict a son's IQ to be if his father has an IQ of 120? What if the father had an IQ of 140?

f) If you knew that the original data included fathers with IQs from 108 to 145, explain why it would be inappropriate to use your model to predict a son's IQ if his father's IQ were 170.

6.4 More Least-Squares Regression

Learning Objectives

- Construct scatterplots using technology
- Calculate and graph the least-squares regression line using technology
- Calculate the correlation coefficient using technology
- Use the LSRL to make predictions
- Understand interpolation and extrapolation
- Interpret the slope and the y-intercept of the LSRL

Multimedia Links

For an introduction to what a least squares regression line represents, see bionicturtled.com, Introduction to Linear R (5:15).

http://www.youtube.com/watch?v=ocGEhiLwDVc

For an applet that will calculate correlation and the least squares regression line, see

http://illuminations.nctm.org/lessonDetail.aspx?ID=L456

Section 6.4 Exercises

1) Mr. Exercise wanted to know whether or not customers continued to use their equipment after they purchased it. He contacted an SRS of his customers who had purchased an exercise machine during the past 18 months. His findings are summarized in the following table. We began to look at his data in section 6.1. We are now going to analyze it further.

# months owned machine	# hours exercise per week
1	8
5	4.5
7	3
4	6
9	2
14	1.5
5	7
11	4
3	6.5
6	4

a) Construct a scatterplot. Calculate the LSRL and add it to your graph. Sketch your graph and report your equation. Be sure to identify your variables.

b) Identify and interpret the slope in the context of the problem.

c) Identify and interpret the y-intercept in the context of the problem.

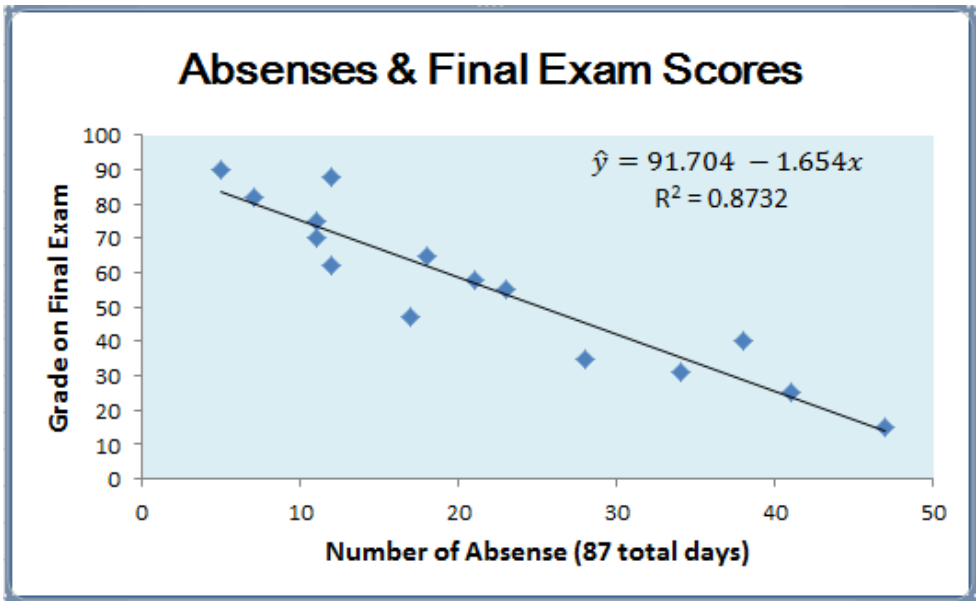
d) What is the correlation coefficient? What are the two things that this statistic tells about the relationship between these two variables?

e) Based on your model, how many hours would you predict a person who has owned the machine for 12 months to exercise? 5 months?

f) Based on your model, if a person claims to exercise 9 hours per week, how long would you suspect that they had owned the machine?

2) A college professor was becoming annoyed by how many of his students were absent during his 8:00 a.m. section of Philosophy 103. He decided to analyze whether these absences were affecting students

learning the material or not. He assigned his TA the task of keeping track of attendance. At the end of the semester he compared students' grade on the final exam with the number of times they had been absent. His findings are displayed in the following graph.



a) Identify the explanatory and response variables.

b) Describe the relationship between these two variables (S.C.O.F.D).

c) Jeremy was absent 25 times. What would you predict his score on the final exam to be? Lucy overslept and missed 43 classes. What would you predict for her score on the final?

d) Calculate the correlation coefficient (r). What two things does this statistic tell you about the association between these two variables? *(Hint: you were given R^2)*

e) Interpret the meaning of -1.654 in the context of this problem.

3) The following table shows the grade level and reading level for 5 students. Treat grade level as the explanatory variable as you do the following.

Grade vs. Reading

Student	Grade	Reading Level
A	2	7
B	6	14
C	5	12
D	4	9
E	1	4

a) Use your calculator to create a scatterplot. Then calculate the LSRL and the correlation coefficient for this data. Report your findings.

What if it was found that student E was actually in grade 8? How would this affect the LSRL and/or the correlation?

Grade vs. Reading

Student	Grade	Reading Level
A	2	7
B	6	14
C	5	12
D	4	9
E	8	4

b) Use your calculator to create a new scatterplot. Then calculate the LSRL and the correlation coefficient for the changed data. Report your findings.

c) What changes do you notice between your answers to (a) and (b)? Explain why these changes occurred.

4) The table below shows the nutritional information for Taco Bell Burritos as reported on the website: http://www.tacobell.com. Choose two of the variables to analyze (avoid using trans fat & sugars).

item	serving size (g)	calories	calories from fat	saturated fat (g)	total fat (g)	trans fat (g)	cholesterol (mg)	sodium (mg)	carbohydrates (g)	dietary fiber (g)
Burritos										
1/2 lb.* Cheesy Potato Burrito	248	540	230	7	26	0.5	45	1360	59	7
1/2 lb.* Combo Burrito	241	460	160	7	18	0.5	45	1330	53	9
7-Layer Burrito	283	500	160	6	18	0	20	1090	69	1
Bean Burrito	198	370	90	3.5	10	0	5	980	56	1
Beefy 5-Layer Burrito	245	540	190	8	22	0	35	1280	68	9
Beefy Nacho Burrito	186	470	180	6	20	0	30	990	58	4
Burrito Supreme® - Chicken	248	400	110	5	12	0	40	1060	51	7
Burrito Supreme® - Steak	248	390	110	5	13	0	30	1100	51	7
Burrito Supreme® – Beef	248	420	140	6	16	0	35	1100	53	9
Chili Cheese Burrito	156	380	150	8	17	0.5	35	930	41	5
Fresco Bean Burrito	213	350	70	2.5	8	0	0	990	57	1
Grilled Chicken Burrito	177	430	170	5	18	0	35	870	48	3
XXL Grilled Stuft Burrito - Beef	445	880	370	14	42	1	75	2050	95	1
XXL Grilled Stuft Burrito - Chicken	445	840	310	11	35	0	85	1970	92	1
XXL Grilled Stuft Burrito - Steak	445	820	320	12	36	0.5	70	2050	92	1

a) What will you be using as your explanatory and response variables?

b) Construct a scatterplot. Label your axes.

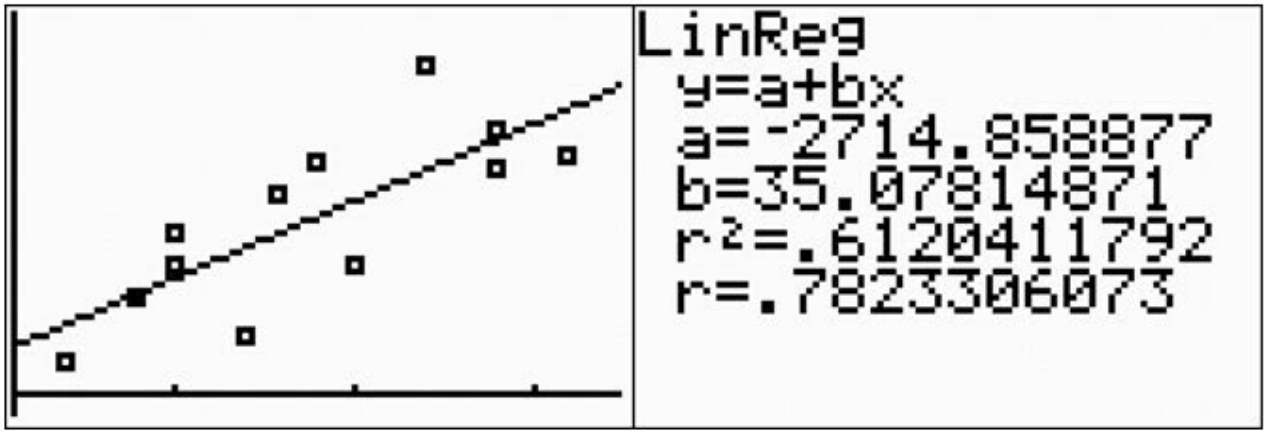
c) Describe the association (S.C.O.F.D.).

d) Calculate the LSRL and the correlation. Report them. Be sure to define your variables. Add the line to your graph in part (b).

e) Use your model to make a prediction that involves interpolation.

f) Use your model to make a prediction that involves extrapolation.

5) *Interpret the calculator output.* The lifeguard at the Swimtastic Pool & Water-Slides decided to keep track of how many people came to the pool each day and compare this to the predicted high temperature for that day. The temperatures ranged from 82° to 96° during his data collection time period. He used the number of people as the response variable. Use this scatterplot and regression output from a TI-84 plus to answer the questions that follow.



a) Write the regression equation. Define your variables.

b) Identify and interpret the slope in the context of the problem.

c) What are the two things that the correlation tells us in this situation?

d) Based on this model, how many people would you predict on a 91° day? How about a 45° day? Are both of these predictions reasonable? Why or why not?

e) What does r tell us in this situation?

6.5 Chapter 6 Review

In this chapter, we have learned that when working with bivariate, numerical data it is important to first identify whether there is an explanatory and response relationship between the two variables. Often one of the variables, the explanatory (independent) variable, can be identified as having an impact on the value of the other variable, the response (dependent) variable. The explanatory variable should be placed on the horizontal axis, and the response variable should be placed on the vertical axis. Next we learned how to construct a visual representation, in the form of a scatterplot, so that we can see what the association looks like. A scatterplot helps us see what, if any, association there is between the two variables. If there is an association between the two variables, it can be identified as being strong if the points form a very distinct form or pattern, or weak if the points appear more randomly scattered. If the values of the response variable generally increase as the values of the explanatory variable also increase, the data has a positive association. If the response variable generally decreases as the explanatory variable increases, the data has a negative association. We also are able to see the form of the pattern, if any, in the graph.

When the data looks reasonably linear, we learned how to use technology to calculate the least-squares regression line and the correlation coefficient. The least-squares regression line is often useful for making predictions for linear data. However, we now know to beware of extrapolating beyond the range of our actual data. Correlation is a measure of the linear relationship between two variables – it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other things may be causing the two correlated variables to relate as they do. We learned how to interpret the linear correlation coefficient and that it can be greatly affected by outliers and influential points. Also, just because two variables have a high correlation, does not mean that they have a cause-and-effect relationship. Correlation ≠ Causation!

Beyond constructing graphs and calculating statistics, we learned how to describe the relationship between the two variables in context. The acronym we learned to help us remember what to include in our descriptions is *S.C.O.F.D.* This tells us to describe the strength of the association, to be sure that our description is in context, to mention any outliers or influential points that we observe, and to describe the form and the direction of the relationship. We also learned how to interpret the slope and y-intercept of the least-squares regression line in context. Even though we are doing easy calculations, statistics is never about meaningless arithmetic and we should always be thinking about what a particular statistical measure means in the real context of the data.

Chapter 6 Review Exercises

Answer the following as TRUE or FALSE.

1) A negative relationship between two variables means that for the most part, as the x variable increases, the y variable increases.

2) A correlation of -1 implies a perfect linear relationship between the variables.

3) The equation of the regression line used in statistics is $\hat{y} = a + bx$

4) When the correlation is high, one can assume that x causes y.

Complete the following statements with the best answer.

5) The symbol for the Correlation coefficient is _____

6) A statistical graph of two variables is called a(n) _____.

7) The _____ variable is plotted along the x-axis.

8) The range of r is from _____ to _____.

9) The sign of r and _____ will always be the same.

10) LSRL stands for _____-_____.

11) If all the points fall on a straight line, the value of r will be _____ or _____.

12) If r = -0.86, then r^2 = _____.

13) If r^2 = 0.77, then r = _____ or _____.

14) Using an LSRL to make predictions outside the range of our original data is called _____-_____.

15) Using an LSRL to make predictions within the range of our original data is called _____-_____.

16) When describing the relationship visible in a scatterplot, the acronym S.C.O.F.D. stands for _____-_____-_.

17) Suppose that a scatterplot shows a strong, linear, positive relationship, and the correlation coefficient is very high. However, both of the variables are actually increasing due to some outside lurking variable. This relationship suffers from _____.

18) Suggest possible lurking variables to explain the high correlations between the following variables. Consider whether common response, confounding, or coincidence may be involved.

a) The number of cell phones being made has been increasing over the past 15 years. So has the number of starving children. Do cell phones cause starvation?

b) The stress level of all of the employees at a certain company has been going up consistently over the past year. During this time, they have received three pay bumps. Does this mean that higher pay is causing the stress?

c) Suppose that a study shows that the number of hours of sleep a person gets is negatively correlated the number of cigarettes a person smokes. Does this mean that not sleeping causes a person to smoke more cigarettes?

19) Some researchers wanted to determine how well the number of beers consumed can predict what a person's blood alcohol content will be after a given length of time. They set up an experiment in which several volunteers each drank a randomly selected number of beers during a given time period. The volunteers were between 21 and 25 years of age, but all ranged in gender and in weight. Exactly three hours after they began to drink the beers, their BAC level was measured three times. The three measurements were averaged and the results are given in the following table. *(This is fictitious data, but is based on calculations from the BAC calculator at: http://www.dot.wisconsin.gov)*

Number of Beers Consumed (3 hours)	10	2	4	6	8	3	3	7	8	5	9	4	6	2
BAC Level	0.29	0.034	0.094	0.1	0.135	0.025	0.062	0.23	0.225	0.127	0.137	0.13	0.06	0.01

a) Identify the explanatory and response variables and construct a scatter-plot (be neat & label your axes).

b) Calculate the LSRL and correlation. Report the equation and add it to your scatter-plot? Identify your variables *(report what x and y stand for).*

c) Identify and interpret the slope in context.

d) Identify and interpret the y-intercept in context.

e) If a person drinks 6 beers during this time period, on average what do you predict the person's BAC will be?

f) If a person drinks 15 beers during this time period, on average what do you predict the person's BAC will be?

g) Are you confident in both of the previous answers? Why or why not?

20) When investigating car crashes, it is often necessary to try to determine the speed at which a vehicle was traveling at the time of the accident. Investigators are able to do this by measuring the length of the skid mark left by the vehicle in question. The following table lists several speeds (mph) based on the skid length (feet), according to the Forensic Dynamics website: http://forensicdynamics.com.

SPEED BASED ON SKID LENGTH

Skid Length (feet)	Estimated Speed (mph)
45	30.68
20	20.45
56	34.23
8	12.93
78	40.4
93	44.11
165	58.75
115	49.05
142	54.51
184	62.05
215	67.07
247	71.89

a) Identify the explanatory and response variables and construct a scatter-plot (be neat & label your axes).

b) Calculate the LSRL and add it to your scatter-plot? Report the equation and identify your variables.

c) Describe the relationship you see in the scatter-plot (S.C.O.F.D.). *Be thorough & use complete sentences!* Be sure that you explain the relationship in the context of the problem (overall trend between the two variables).

d) What is the correlation coefficient? Based on your scatterplot and the value of r, how well do you feel that your model fits this data? Explain

e) What is the predicted speed if the skid mark is 157 feet? If it were 36 feet?

f) Would you expect predictions beyond 250 feet to generally over-estimate or under-estimate the actual speed of the vehicle? Why?

Chapter 7

The Normal Distribution

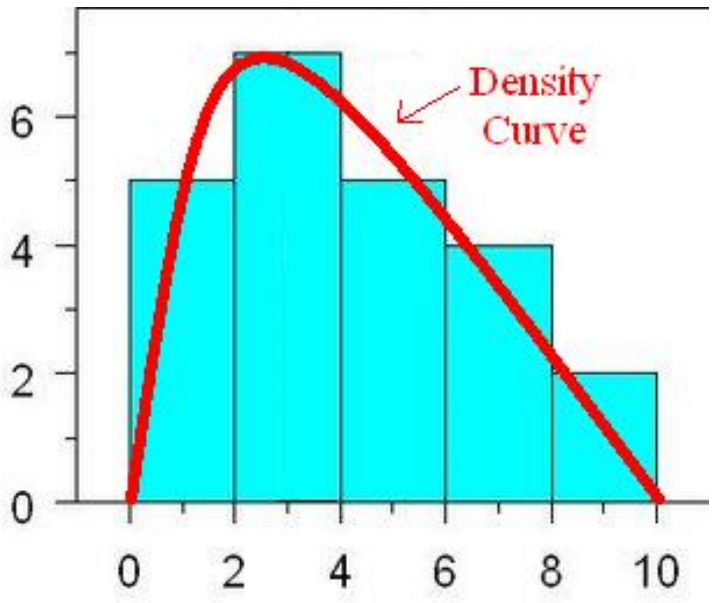
7.1 Introduction to the Normal Curve



Learning Objectives

- Understand how a density curve can be used to approximate the data in a histogram
- Understand how to visually identify the mean and standard deviation of a normal distribution
- Be able to tie the concepts of percentages in the 68-95-99.7 empirical rule to normal distributions

In previous chapters we have seen how data can be represented by histograms. A **density curve** is a curve that gives an approximate description of a distribution. The curve is smooth, so any small irregularities in the data are ignored. A density curve for a particular histogram is shown below. Perhaps the most important thought to remember about a density curve is that it represents 100% of the data. In other words, the area under any density curve is equal to 1. This is important because it allows us to ask probability questions about a population. For example, we might ask how likely is it that a teenager has a shoe size of 8 or larger.

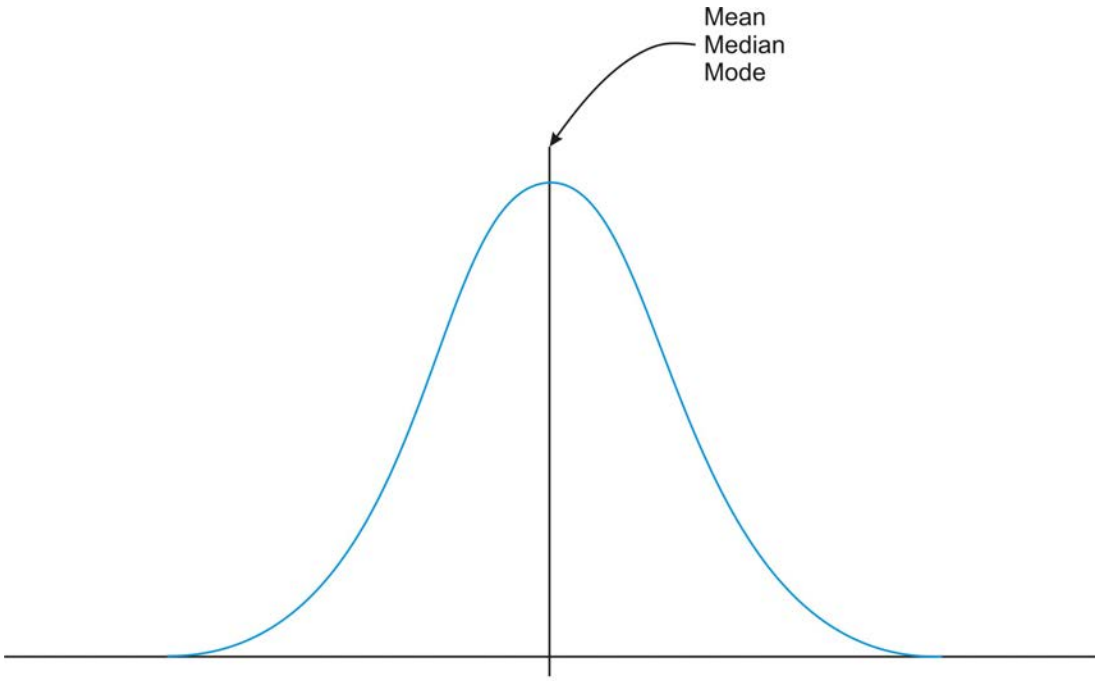


In our chapter, we will focus on a special density curve called the normal curve. Have you ever wondered if you are 'normal'? You probably are normal in most ways, but there may be some things about you that might not be considered normal by the mathematical definition. If you are on the high school baseball team, do you throw the baseball at a 'normal' speed? Is your hair a 'normal' length? Do you drive at a 'normal' speed on the freeway? Our goal this chapter is to gain an understanding of what 'normal' really is and how to properly calculate within the Normal Distribution. We have seen skewed distributions before. The density curves in the following figure show one density curve that is skewed left and one that is skewed right.





A **normal curve** is neither skewed left nor right and is often referred to as 'the bell curve' because of its shape. It is symmetrical. In addition, as you get closer and closer to the middle of the curve, there is a higher frequency of results. The **mean** (along with the median and mode) always lands at the center of a normal distribution. When dealing with the mean in previous chapters, we have used the symbol $\overline{x}$ because the data came from a sample. Normal distributions deal with an entire population instead of just a sample and we will use the symbol μ (Greek letter mu) to mark the mean of a normal distribution for an entire population. The mean is one of two key values needed to make a proper sketch and analysis of a normal distribution. The curve shown below represents a normal distribution and is a good representation of what a normal curve looks like.



Normal Distribution Center

Note that the amount of data to the right of the mean is the same as the amount of data to the left of the mean. Thinking about the definition of the median, this suggests that the mean and median are located at the same point. The other key component used to construct and analyze a normal distribution is the **standard deviation**. The standard deviation is a measure of spread and can be loosely thought of as a 'typical' distance from the mean. You may have calculated the standard deviation before for data sets either by hand or by using your calculator and looked for the S_x in the statistical calculations summary screen. The symbol S_x is used for the standard deviation whenever data is collected through the use of a sample from a population. When dealing with the normal distribution, we will use the symbol σ (Greek letter sigma) to represent the standard deviation. The σ symbol indicates that the standard deviation of the entire population is known. Visually, the standard deviation can be seen as the distance from the mean to an **inflection point**. An inflection point is located on a curve at the point where the curve changes from **concave up** (bent up) to **concave down** (bent down) or vice versa. On the normal curve in Figure 7.1, the mean is 23 and the standard deviation is 3.

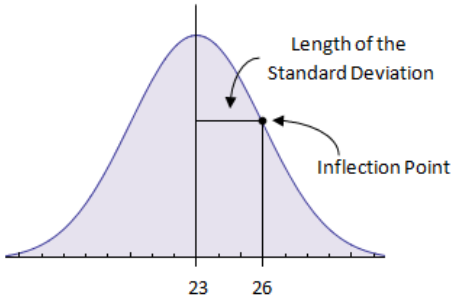
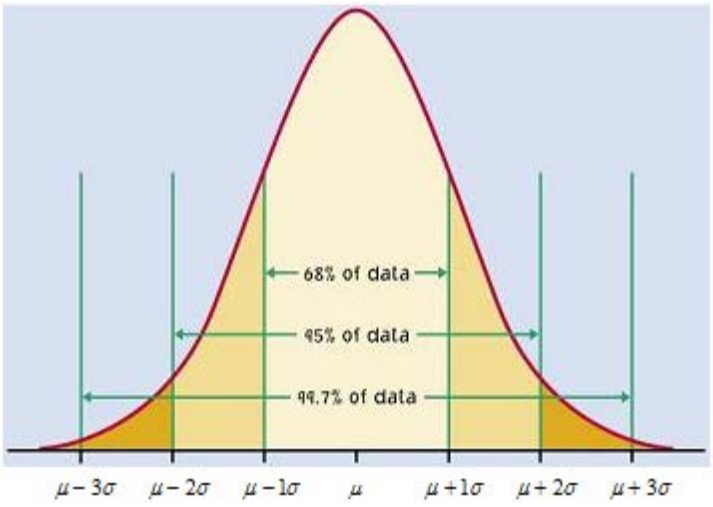


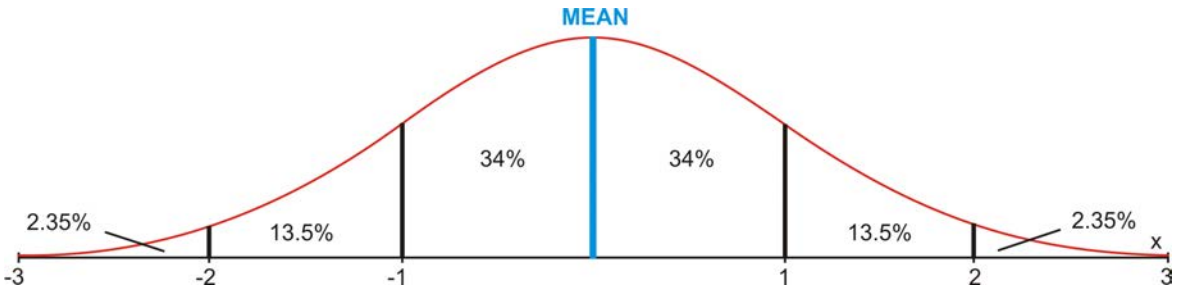
Figure 7.1

The 68-95-99.7 Rule

It is now time to make use of the some of the special characteristics of the normal curve. As mentioned earlier, 100% of all results fall somewhere under the normal curve. It turns out that approximately 68% of all results are within one standard deviation of the mean, 95% of all results are within 2 standard deviations of the mean, and 99.7% of all results land within three standard deviations of the mean. These percentages are illustrated in the graphic below.



The numbers on the bottom represent the number of standard deviations from the mean. For example, the $\mu - 1\sigma$ marks the point one standard deviation below the mean. Some simple addition and subtraction allows us to be very specific in the percents of the data that land in the sections of the normal curve as shown below.



Can you see the 68-95-99.7 rule here?

Example 1

Suppose the mathematics portion of the SAT exam is normally distributed with a mean of 500 and a standard deviation of 100.

a) Sketch a normal curve for this situation marking the mean and the values 1, 2, and 3 standard deviations above and below the mean.

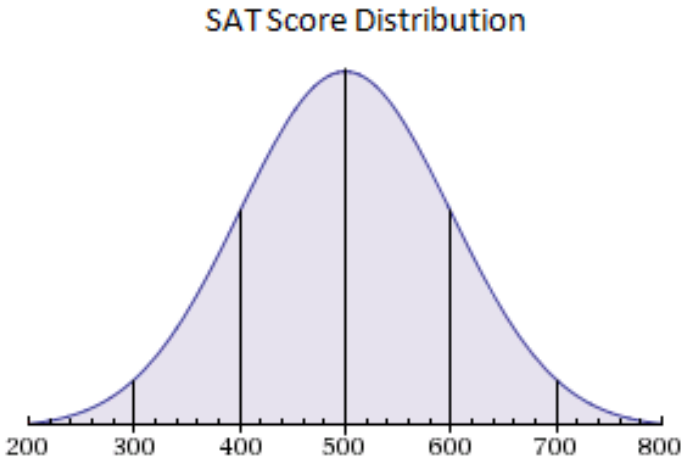
b) Using the 68-95-99.7 rule, approximately what percent of students scored at least 600 on this test?

c) Between approximately which two scores did the middle 95% of students score?

d) Suppose that 4600 students take the exam this month. How many of those students should we expect to obtain a score of at least 700?

Solution

a)



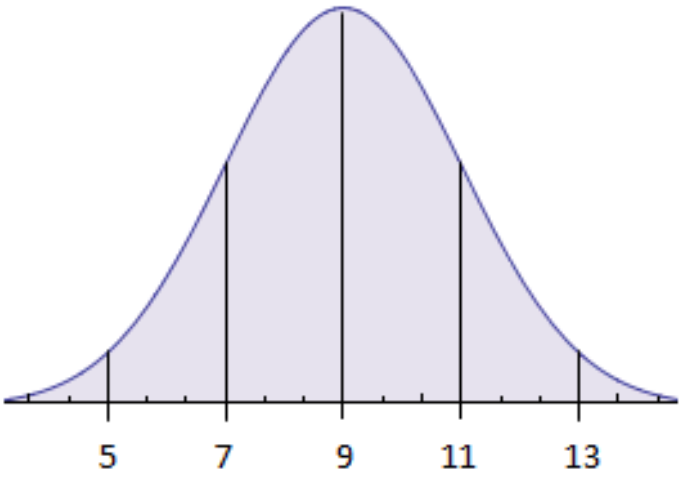
b) We know that 50% of all results are below the 500 marker and that 34% of all results land between 500 and 600. We have used up 50% + 34% = 84% of all results. This tells us that 100% - 84% = 16% of all students scored above 600 on the mathematics portion of the SAT.

c) The middle 95% of all students scored within 2 standard deviations of the mean or between 300 and 700.

d) A score of 700 marks the boundary two standard deviations above the mean such that only 2.5% of all test takers will score at least 700. 2.5% of 4600 is 115 students.

Example 2

The normal curve below represents the number of races that a typical racehorse will run in one calendar year.



a) Approximately what percent of racehorses will run between 5 and 11 races during a calendar year?

b) What are the values of the mean and standard deviation for the distribution shown?

Solution

a) Add 13.5% + 34% + 34% to get 81.5% so 81.5% of racehorses run between 5 and 11 races per year.

b) The mean racehorse will run 9 races per year with a standard deviation of 2 races.

What is Normal?

Let's now go back and try to think about our original question "What is normal?" In mathematics, the middle 95% is often (but not always) considered our 'normal' group. For example, suppose the ACT exam is normally distributed with a mean of 18 and a standard deviation of 6. Our 'normal' group would be comprised of those students who scored anywhere within two standard deviations of the mean or from 6 to 30 on the exam. A student who scored 31 or higher on the exam would have achieved an exceptional score. We might say that this student was not normal with regards to their ACT score.

Normal distributions are not as common as you might think. What if we measured the lengths of shoes of teenagers? Many students think that this would be normal when in fact, there are a couple of contributing factors that might tip us off that the situation may not be normal. First of all, teenagers encompass a large population. Most of those who are in their upper teen years have finished growing into their adult shoe size length whereas many of the younger teens are still growing. This would tend to give us a slightly larger percentage of smaller shoe lengths than we might expect from a normal distribution. In addition, teenagers include males and females. This may lead to us seeing a situation which might be bi-modal. We might expect to see a peak at the most common male lengths and at the most common female lengths.

Example 3

Which situation below is most likely to produce a normal distribution?

a) The heights of all adults.

b) The wingspans of three year-old American eagles.

c) The number of teeth that Americans adults have.

Solution

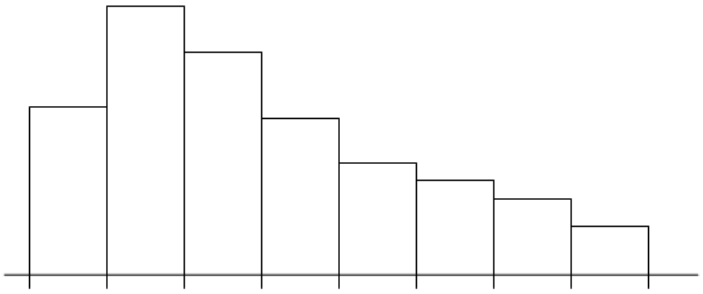
The correct answer is b). Three year-old American eagles have an average wingspan and we would expect that there are quite a few eagles at that wingspan or very close to it. As we move further and further up and down from that average, we would expect to see fewer and fewer eagles with those wingspans. Answer a) could be ruled out quickly in that the heights here do not specify a particular group. For example, this data would include males and females. Answer c) is out because the vast majority of American adults have 32 teeth. As we move away from 32, there are some people with fewer teeth due to a variety of reasons but there are virtually no people with more than 32 teeth. We should see symmetrical results if this was a normal distribution.

http://www.anoka.k12.mn.us/education/page/download.php?fileinfo=Ny0xX0ludHJvX3RvX05vcm1hbF9DdXJ2ZV9

Problem Set 7.1

Exercises

1) Consider the histogram shown below.



a) Make a sketch of the histogram and overlay a sketch of a density curve for the histogram.

b) What is the area under your density curve?

c) What is the shape of the density curve?

2) A roadside bait salesman digs up worms to sell to fishermen. It turns out that the worms have a mean length of $\mu = 112$ mm and a standard deviation of $\sigma = 12$ mm.

a) Draw and label a normal curve for this distribution. Include lines for the mean and for 1, 2, and 3 standard deviations above and below the mean.

b) What percentage of the worms will have lengths longer than 112 mm?

c) What percentage of the worms will have lengths between 100 and 124 mm long?

d) What percentage of the worms will have lengths between 100 and 112 mm long?

e) What percentage of the worms are longer than 124 mm?

f) What percentage of the worms are shorter than 88 mm?



3) Sketch a normal curve which has a mean of 13 pounds and a standard deviation of 3 pounds. Include lines for the mean and for 1, 2, and 3 standard deviations above and below the mean.

4) Not all 12-ounce cans of soda are the same. It turns out that the average 12-ounce can of soda does contain twelve ounces of soda, but the amount of soda is normally distributed with a standard deviation of 0.15 ounces. Fill in the blanks for each statement below.

a) The middle 68% of all 12-ounce soda cans contain between _____ & _____ ounces of soda.

b) The middle 95% of all 12-ounce soda cans contain between _____ & _____ ounces of soda.

c) The middle 99.7% of all 12-ounce soda cans contain between _____ & _____ ounces of soda.

5) Figure 7.2 shown below shows an approximate distribution of the number of fish caught by the competitors during a one hour pan-fishing contest. Give the approximate values of the mean and the standard deviation for the distribution.

6) Suppose the weights of adult males of a particular species of whale are distributed normally with a mean of 11,600 pounds and a standard deviation of 640 pounds.

a) Draw a normal curve for this situation. Use vertical lines to mark and label the mean and 1, 2, and 3 standard deviations above and below the mean.

b) What percent of these whales weigh less than 10,320 pounds?

c) Between what two weights do the middle 99.7% of these whales weigh?

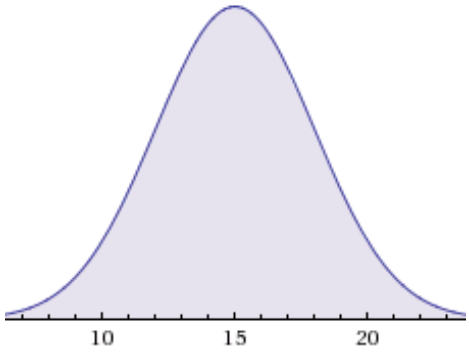


Figure 7.2

d) What percent of these whales weigh between 10,320 pounds and 12,240 pounds?

7) Which situation is most likely to be normally distributed? Explain your reasoning.

i) The hair lengths for all the Statistics and Probability students who have Mr. Johnson as a teacher.

ii) The prices of all new Ipod Touches that are sold in Minnesota this week.

iii) The average running times for all 4th grade boys at Andover Elementary in the 50 yard dash.

8) Suppose a standard incandescent light bulb will run an average of 400 hours before burning out. Of course, some bulbs burn out sooner and some last longer. Suppose that the average lives of these bulbs is normally distributed with a standard deviation of 35 hours.

a) Sketch and label a normal curve to illustrate this situation.

b) What percent of these bulbs will burn out in 400 hours or less?

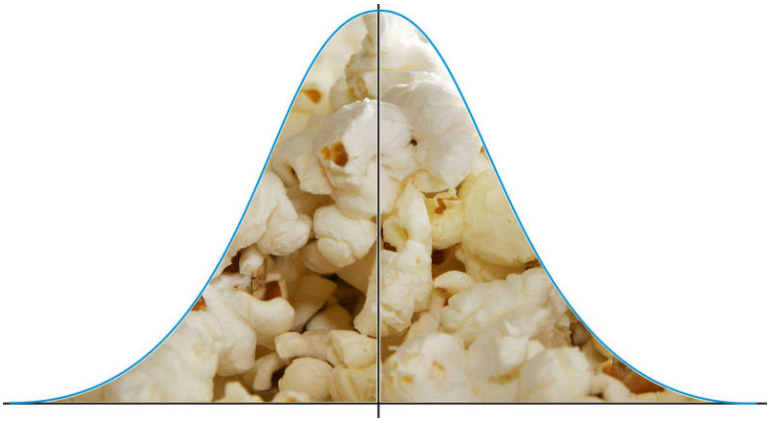
c) If you are lucky, your bulb will last longer than advertised. What percent of bulbs should last 435 hours or more? What percent of bulbs will last 470 hours or more?

d) If you had 5000 bulbs that you needed for use in a large office building, how many would you expect to last at least 365 hours?

9) Suppose that the time that it takes for a popcorn kernel to pop produces a normal distribution with a mean of 145 seconds and a standard deviation of 13 seconds for a standard microwave oven.

a) It is usually not a good idea to let the microwave oven run until all the kernels are popped because some of the popcorn will start to burn. Suppose the ideal time to shut off the microwave oven is after about 97.5% of the kernels have popped. When will 97.5% of the kernels be popped?

b) Between what two times will we see the middle 68% of kernels popped?



10) After a great deal of surveying, it is determined that the average wait times in the cafeteria line are normally distributed with a mean of 7 minutes and a standard deviation of 2 minutes. Suppose that 400 students are released to the cafeteria for 2nd lunch.

a) Approximately how many students will have to wait more than 5 minutes for their food?

b) Approximately how many students will have to wait more than 11 minutes for their food?

11) Sudoku is a popular logic game of number combinations. It originated in the late 1800s in the French press, *Le Siècle*. The mean time it takes the average 11th grader to complete the Sudoku puzzle below was found to be 19.2 minutes, with a standard deviation of 3.1 minutes.

a) Draw a normal distribution curve to represent this data.

b) Suppose Andover High School is going to put together a Sudoku team. The coach has decided that she will only consider players who score in the fastest 2.5% of the junior class as she puts together the team. How fast must a student solve a puzzle to be in the top 2.5% of puzzle solvers?

c) If there are 400 kids in the Andover junior class, how many of them will be able to solve the Sudoku puzzle below in 16.1 minutes or less?

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

12) In order to qualify for undercover detective training, a police officer must take a stress tolerance test. Scores on this test are normally distributed with a mean of 60 and a standard deviation of 10. Only the top 16% of police officers score high enough on the test to qualify for the detective training. What is the cutoff score that marks the top 16% of all scores?

Review Exercises

13) Use your calculator to find the mean and standard deviation of the data set below.

3, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8

14) A pet store must select 2 dogs and 2 cats for display in their front window. In how many ways can this be done if there are 16 dogs and 12 cats available to choose from?



15) By hand, give the five number summary for the data set below.

3, 5, 5, 6, 8, 9, 10, 10, 12, 13, 13, 13, 14, 15, 17, 19, 19, 20

16) A student conducts a survey in which 100 tenth-graders are asked "What is your favorite item on the lunch menu at school today?" The student decides to conduct this survey by handing each tenth-grader a survey sheet while they are eating and asking them to fill it out and turn it in to room P202 by the end of the day. Why will this survey method have a problem with bias?

7.2 Z-Scores, Percentiles, and Normal CDF

Learning Objectives

- Be able to calculate and understand z-scores
- Understand the concept of a percentile and be able to calculate it for a particular result
- Be able to calculate percentages of data above, below, or in between any specific values in a normal distribution

- Be able to use z-scores to compare results for two different but related situations

In section 7.1, we analyzed normal distributions and specific situations in which analysis was done for data which followed the 68-95-99.7 rule exactly. The truth of the matter is that most situations require us to answer questions that do not reference exact whole numbers of standard deviations above or below the mean. What if we asked a student what their actual score would be if they were in the top 10% of ACT test takers? We need a tool to help us deal with these types of situations.

Our first tool will be the z-score formula. The **z-score** is a measure of how many standard deviations above or below the mean a particular value is. If a z-score is negative, the result is below the mean and if it is positive, the result is above the mean. For example, if the ACT mathematics exam scores are normally distributed with a mean of 18 and a standard deviation of 6, then an ACT score of 30 would be equivalent to a z-score of 2 because 30 would be 2 standard deviations above the mean.

The formula below gives a quick way to calculate z-scores. In the formula, 'x' is the observation, μ is the mean of the distribution, and σ is the standard deviation for the distribution.

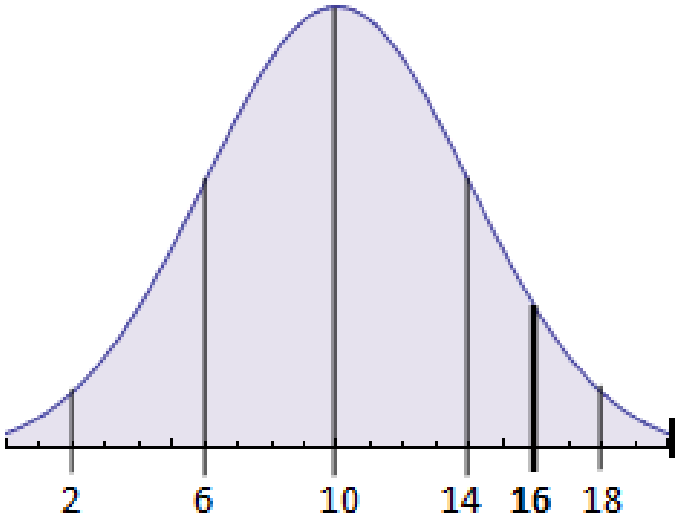
$$z = \frac{x - \mu}{\sigma}$$

Example 1

Suppose the mean length of the hair of 10th grade girls is 10 inches with a standard deviation of 4 inches. What would be the z-score for hair length for a 10th grade girl whose hair is 16 inches long and what does it mean in terms of the normal curve?

Solution

It is often a good idea to draw a sketch for these sorts of situations so we can visualize what is happening.



Because 16 is located between 1 and 2 standard deviations above the mean, we expect a z-score between 1 and 2. Use the formula $z = \frac{x-\mu}{\sigma}$ to calculate the z-score. Our observation, x, is 16 inches while the mean is $\mu = 10$ inches and the standard deviation is $\sigma = 4$ inches. $z = \frac{16-10}{4}$ or z = 1.5. This tells us that a hair length of 16 inches will be 1.5 standard deviations above the mean.

Example 2

Suppose that the z-score for a particular 10th grade girl's hair length is z = -1.25. What is the length of the girl's hair?

Solution

We will use the z-score formula to find our answer.

$-1.25 = \frac{x-10}{4}$

$-5 = x - 10$

$5 = x$

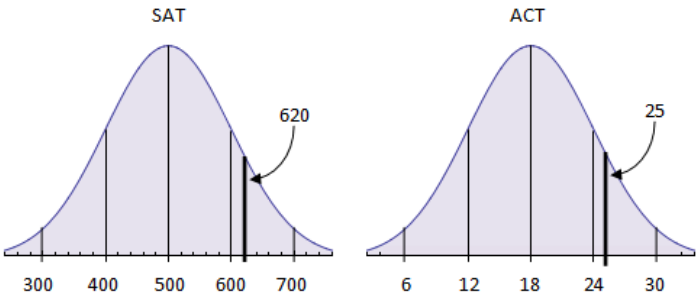
The length of the hair for this girl would be 5 inches.

Example 3

Suppose a student can either submit only their SAT score or their ACT score to a particular college. Suppose their SAT score was 620 and that the SAT has a mean of 500 and a standard deviation of 100. Suppose also that the same student scored a 25 of their ACT exam and that the ACT exam has a mean of 18 and a standard deviation of 6. Which score should the student submit?

Solution

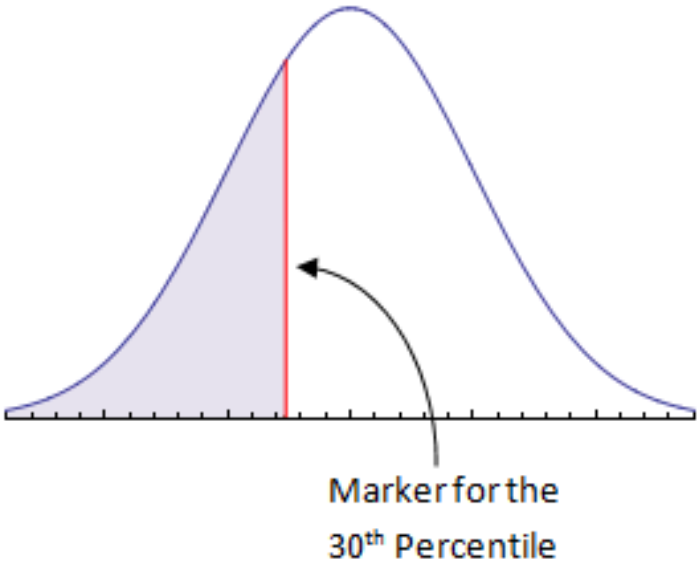
Looking at the diagram below, it is not exactly clear which score is better. They appear to be quite similar and we will need to do some calculations to make a distinction.



Calculate the z-score for each exam. For the SAT, $z = \frac{620-500}{100} = 1.2$. For the ACT, $z = \frac{25-18}{6} \approx 1.17$. Since the z-score is higher on the SAT, the student should submit the SAT exam score.

Percentiles

In order to understand how to apply z-scores beyond what we have already done, we must first understand percentiles. A **percentile** is a marker on a normal curve such that the marker is greater than or equal to that percentage of results. For example, suppose you are at the 30th percentile for how fast you type. This means that you can type faster than 30% of all people. The percentile can also be thought of as the percent of area to the left of its marker. The graphic below shows where the 30th percentile is located. The shaded area to the left of the marker represents 30% of the normal curve.



It is very common for colleges and universities to use percentiles for entrance criteria. For example, a rather elite university might require that you score at the 90th percentile or higher on your ACT exam to be considered for admissions. Doctors often use percentiles to track the growth of babies. For example, can you picture what a baby would look like that is at the 70th percentile for weight and the 25th percentile for length?

Now we must ask what percentiles have to do with z-scores. Find the **Normal Distribution Table in Appendix A, Part 2** of your book. Let's examine the z-score of -1.25 from Example 2. Find the z-value of -1.2 and then go over until you are under the 0.05 column. A partial table is given in Figure 7.3 below and the value in the cell we are looking for is bold and underlined. The value of 0.1056 can be interpreted as a percentile. This means that the girl in Example 2 has hair that is longer than 10.56% of all girls. In other words, she is at about the 10th or 11th percentile for hair length for 10th grade girls.

Negative z-scores:										
z	**0.09**	**0.08**	**0.07**	**0.06**	**0.05**	**0.04**	**0.03**	**0.02**	**0.01**	**0.0**
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	**0.1056**	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357

Figure 7.3

Example 4

At what percentile for hair length is a 10th grade girl if her hair is 17 inches long?

Solution

Start by determining her z-score which would be $z = \frac{17-10}{4} = \frac{7}{4} = 1.75$. We now go to the Normal Distribution Table in Appendix A, Part 2 of the book. We go across the row with z=1.7 until we are under 0.05. This gives a value of 0.9599. This tells us the girl is at about the 96th percentile for hair length. In other words, this girl's hair is longer than 96% of all 10th grade girls.

'Between' and 'Above' Problems

While it is nice to find percentiles for certain situations, we are often asked for the percentage of results that are between two given parameters or above a given parameter. For example, we might be asked to find the percentage of all 10th grade girls that have hair lengths between 8 inches and 15 inches long. To find these types of results, we often must do multiple z-score calculations and some addition or subtraction.

Example 5

Suppose the weights of adult males of a particular species of whale are distributed normally with a mean of 11,600 pounds and a standard deviation of 640 pounds.

a) What percent of these adult male whales will weigh between 11,000 and 12,000 pounds?

b) What percent of these adult male whales will weigh more than 12,000 pounds?

Solution

a) Begin by finding the z-scores for both of the weights given and get $z = \frac{11{,}000-11{,}600}{640} = \frac{-600}{640} = -0.9375$ and $z = \frac{12{,}000-11{,}600}{640} = \frac{400}{640} = 0.625$. For z=-0.9375, our Normal Distribution Table in Appendix A, Part 2 gives us a value between 0.1736 and 0.1762. Since -0.9375 is closer to -0.94 than -0.93, we will use a value of 0.174. We get a value between 0.7324 and 0.7357 for z=0.625. We will split the difference on this and use 0.734. All that is left to do now is subtract 0.734 and 0.174 to get 0.56 or about 56% of all adult male whales of this species are between 11,000

and 12,000 pounds. The shaded region in the Figure 7.4 below represents about 56% of the normal curve.

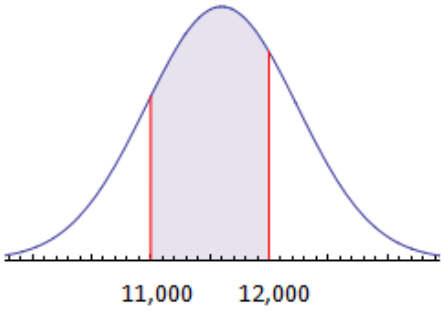


Figure 7.4

b) Use z=0.625 from part a) to get a value from the table of 0.734. This means that 73.4% of all whales weigh 12,000 pounds or less. Therefore, 100%-73.4%=26.6% of all whales weigh more than 12,000 pounds.

Technology

It is also important to note that graphing calculators can be used to quickly solve the types of problems discussed in this section by using the NormalCdf command. Typically, this command requires that four values be entered, the lower bound, the upper bound, the mean, and the standard deviation. In Example 5, we can solve the problem in part a) simply by typing in the command string NormalCdf(11000,12000,11600,640) and obtain the immediate result of 0.5598 or 56%.

Be sure you know how to access this command if you have a graphing calculator. Appendix C has some notes for users of graphing calculators. An online calculator that is very similar to a graphing calculator and gives us the same information can be found at http://wolframalpha.com .

You might also be wondering how to solve a problem using the NormalCdf command if only one parameter is given. Let's revisit Example 4 to see how this works.

Example 6

At what percentile for hair length is a 10th grade girl if her hair is 17 inches long?

Solution

There is only one boundary given in this problem. It is your job to come up with a second boundary. In this case, the percentile we want to calculate is found by finding the percentage of all girls whose hair is 17 inches or less. We will use a lower bound of -100 and an upper bound of 17. We use -100 simply because we are confident that we will not find any results any further left than this. Typically, choose your missing parameter as being so extreme that it will not be even in the realm of possible results. NormalCdf(-100,17,10,4)=0.9599 so the length of the girl's hair is at about the 96th percentile.

Problem Set 7.2

Exercises

For problems 1) through 14) use the following information: On a particular stretch of road, the number of cars per hour produces a normal distribution with a mean of 125 cars per hour and a standard deviation of 40 cars per hour.

1) Sketch a normal curve for this situation. Be sure to label and mark the mean and 1 and 2 standard deviations above and below the mean.

2) What is the z-score for an observation of 165 cars in one hour?

3) What is the z-score for an observation of 85 cars in one hour?

4) Calculate the z-score associated with an observation of 171 cars in one hour.



5) Suppose 135 cars are observed in one hour. At what percentile would this observation occur?

6) Suppose 70 cars are observed in one hour. At what percentile would this observation occur?

7) At what percentile would an observation of 125 cars occur?

8) What is the probability of observing at least 145 cars on the road in a an hour?

9) What is the probability of observing between 100 and 150 cars on the road in an hour?

10) Determine the percentile for an observation of 140 cars on the road in one hour.

11) Determine the percentile for an observation of 65 cars on the road in one hour.

12) Determine the probability of observing between 90 and 130 cars on the road in one hour.

13) Determine the probability of observing at least 160 cars on the road in one hour.

14) Determine the probability of obsering no more than 110 cars on the road in one hour.

For problems 15) through 20) use the following information: The number of ants found in one mature colony of leafcutter ants is normally distributed with a mean of 136 ants and a standard deviation of 14 ants.



15) One ant colony has 165 ants. At what percentile for size is this ant colony?

16) An ant colony has a z-score of -1.35 for size. How many ants would we expect to find in this colony?

17) Another ant colony has 131 ants. What is the z-score for this ant colony?

18) What is the probability of finding an ant colony with 160 ants or less?

19) What is the probability of finding an ant colony with 150 ants or more?

20) What is the probability of finding an ant colony that has between 120 and 155 ants in it?

21) Twin brothers Ricky and Robbie each took a college entrance exam. Ricky took the SAT which had a mean of 1000 with a Standard Deviation of 200 while Robbie took the ACT which had a mean of 18 with a standard deviation of 6. Which brother did better if Ricky scored a 1140 and Robbie scored a 22?

22) Suppose the average height of an adult American male is 69.5 inches with a standard deviation of 2.5 inches and the average height of an adult American female is 64.5 inches with a standard deviation of 2.3 inches. Who would be considered taller when compared to their gender, an adult American male who is 74 inches tall or an adult American female who is 68.5 inches tall? Explain your answer.

23) Professional golfer John Daly is one of the longest hitting golfers in history. Suppose his drives average 315 yards with a standard deviation of 12 yards. Will a drive of 345 yards be in his top 1% of his longest drives? Explain your answer.

Review Exercises

24) What is the area under any density curve equal to?

25) In a standard deck of 52 cards, what is the probability of being dealt two queens if you are dealt two cards from the deck without replacement?



26) In a class competition, each grade (9-12) enters 10 students to run in a 500 meter race. Boys times for 9th graders and 12th graders are given below in seconds. Build a back-to-back stem plot to compare data for the two groups of students.

9th Grade Times = 115, 118, 118, 121, 126, 127, 131, 134, 140

12th Grade Times = 106, 106, 109, 112, 114, 116, 116, 121, 122, 133

27) It turns out that countries that have higher percentages of people with computers also tend to have people who live longer. Is it logical to assume that shipping many computers to countries whose people have lower life-expectancies will help the people in those countries live longer? Answer the question including justification that references either Cause and Effect, Common Response, Confounding, or Coincidence.

28) A sample survey at a local college campus asked 250 students how many textbooks they were currently carrying. The table below shows a summary of the findings. Use the table to determine the expected number of textbooks that an average college student at this campus would be carrying.

Table 7.1: **Textbooks Carried by Students**

# of Books	0	1	2	3
Probability	0.21	0.37	0.32	0.1

7.3 Inverse Normal Calculations



Learning Objectives

- Understand how to use the Normal Distribution Table and the z-score formula to find values for a particular normal distribution given a percentile

- Be able to use the Inverse Normal command on a graphing calculator to find values for a particular normal distribution given a percentile
- Be able to find values for a particular normal distribution given a 'middle' percentage range

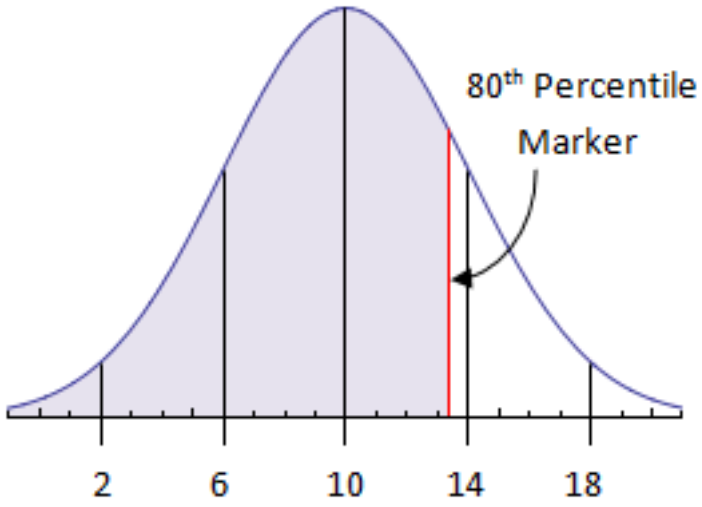
We can now comfortably calculate percentages, percentiles, and probabilities given key information about a normal distribution. It is possible to go the other direction. In other words, if you are told a certain result is at a specific percentile, you can figure out what the actual value is equal to that is at that percentile. The process can be done using the **Normal Distribution Table in Appendix A, Part 2**. Begin by identifying the percentile you are interested in and finding it in the table. From there, put the value from the table into the z-score formula and solve it for the observation in question.

Example 1

Suppose that 10th grade girls have hair lengths that are normally distributed with a mean of 10 inches and a standard deviation of 4 inches. How long would a 10th grade girl's hair have to be in order to be at the 80th percentile for length?

Solution

The figure below shows the distribution of hair lengths and also marks where the 80th percentile is located.



Begin by finding the value closest to .8000 in the Normal Distribution Table. We find our closest value to be .7995 which corresponds to a z-score of 0.84. Put this value into the z-score formula to get $0.84 = \frac{x-10}{4}$.

$0.84 = \frac{x-10}{4}$

$3.36 = x - 10$

$13.36 = x$

A 10th grade girl would have to have a hair length of about 13.4 inches to be at the 80th percentile. This looks to be right based upon comparison to the figure above.

Technology

Once again, it is important to note that technology can be used to solve these types of problems without having to reference the Normal Distribution Table. The command that is commonly used for these types of problems is the Inverse Normal command or InvNorm. The Inverse Normal command requires users to enter the percentile in question, the mean, and the standard deviation. To solve the problem in Example 1, we could have typed in InvNorm(0.80,10,4) and we would have immediately had an answer of 13.366 or about 13.4 inches of hair.

Be sure you know how to access this command if you have a graphing calculator. Appendix C has some notes for users of graphing calculators.An online calculator that can produce the same information can be found at http://wolframalpha.com .

'Middle' and 'Top' Problems

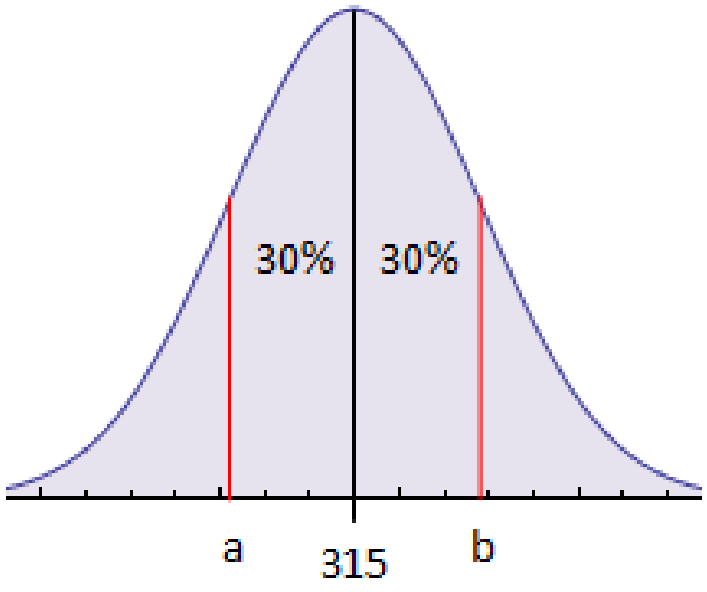
Finally, we are sometimes in situations where we want to know what range of results are found in a middle percentage interval or what value one would have to be at in order to be in a specific top percentage. For example, a car salesman might wish to know what sales prices comprise the middle 50% of his sales to help him learn more about who his customers tend to be or a student might wish to know what they need to score on a test in order to be in the top 10%. Once again, this process can be done with either the Normal Distributions Table or by using technology.

Example 2

Professional golfer John Daly is known for his long drives off the tee. Suppose his drives have a mean distance of 315 yards with a standard deviation of 12 yards. What lengths of drives will constitute the middle 60% of all of his drives?

Solution

The sketch below is helpful in understanding what is happening here.



It is easy to calculate that marker line 'a' is at the 20th percentile and marker line 'b' is at the 80th percentile simply by noting their relationship to the 50th percentile marker. In addition, note that 'a' and 'b' clearly enclose the middle 60 percent of all data. From the Normal Distributions Table, we can see that the z-score associated with the 20th percentile is -0.84 and the z-score associated with the 80th percentile is 0.84. We now calculate $-0.84 = \frac{x-335}{14}$ or x = 303.2 yards. A similar calculation at the 80th percentile gives us x = 326.8 yards. In other words, the middle 60% of John Daly's drives will travel between 303.2 yards and 326.8 yards.

We also could have used the Inverse Normal command once we knew the percentiles. InvNorm(.20,335,14) = 323.2 yards and InvNorm(.80,335,14) = 346.8 yards.

Example 3

In a weightlifting competition, the amount that the competitors can lift is normally distributed with a mean of 196 kg and a standard deviation of 11 kg. Only the top 20% of all competitors will be able to advance to the next phase of the competition. What amount must a competitor lift in order to move into the next phase of the competition?

Solution

The key to this problem is noticing that to be in the top 20%, a competitor would actually have to be at the 80th percentile. The z-score at the 80th percentile is z=0.84.

$0.84 = \frac{x-196}{11}$

$9.24 = x - 196$

$205.24 = x$

The competitor would have to lift about 205 or 206 kg. Using a calculator, we get InvNorm(.8,196,11) = 205.26 kg.

Problem Set 7.3

Exercises

1) The Standard Normal Curve is defined as having a mean of 0 and a standard deviation of 1.

a) What is the z-score associated with a result at the 84th percentile?

b) What is the z-score associated with a result at the 16th percentile?

c) Find a z-score such that only 5% of the Standard Normal Curve is to the right of that z-score.

d) Find a z-score such that only 35% of the Standard Normal Curve is the left of that z-score.

e) Find the two z-scores such that the middle 50% of the Standard Normal Curve is between the two z-scores.

2) Doctors often monitor their patients blood-glucose levels. It is known that for blood-glucose levels, $\mu = 85$ and $\sigma = 25$.



a) Draw and label sketch of the normal distribution for this situation marking the mean and 1, 2, and 3 standard deviations above and below the mean.

b) It turns out that doctors consider the blood-glucose level of a patient to be normal if the level is in the middle 94% of all results. What range of blood-glucose levels constitute the middle 94% of all results?

c) Patients are considered to be high risk for diabetes if their blood-glucose test comes back in the top 1% of all results. What blood-glucose level marks the start of the top 1% of blood-glucose levels?

d) Doctors also show concern if there is too little blood-glucose in a patient's system. They will prescribe treatments to patients if their blood-glucose is in the lowest 2% of all patients. What is the blood-glucose level that marks this boundary?

3) For a given population of high school juniors and seniors, the SAT math scores are normally distributed with a mean of 500 and a standard deviation of 100. For that same population, the ACT math exam has a mean of 18 with a standard deviation of 6.

a) One school requires that students score in the top 10% on their SAT math exam for admission. What is the minimum score that a student must achieve to be considered for this school?

b) Another school requires that students score in the top 40% on their ACT math exam for admission. What is the minimum score that a student must achieve to be considered for this school?

c) One particular school likes to focus on mid-level students and so they only accept students who are in the middle 50% of all ACT math test takers. Between what two scores must a student achieve in order to be considered for acceptance into this school?

d) One student boasts that they scored at the 85th percentile on their ACT math exam. Another student brags that they scored a 620 on the SAT math exam. Who did better?

4) Many athletes train to try to be selected for the US Olympic team. Suppose for the men's 100 meters, the athletes being considered for the team have a mean time of 10.06 seconds with a standard deviation of 0.07 seconds. In the final qualifying event for the team, only the top 20% of runners will be selected. What time must a runner get to be in the top 20%?



5) A high school basketball coach notices that taller players tend to have more success on his team. As a result, the coach decides that only the tallest 25% of the boys in the 11th and 12th grades will be allowed to try out for the team this year. Suppose that the mean height of 11th and 12th grade boys is 5 feet 9 inches with a standard deviation of 2.5 inches. How tall must a player be in order to be able to try out for the team?

6) A student comes home to his parents and excitedly claims that he is in the top 90% of his class. Explain why this might not be worth getting excited about.

7) At a certain fast-food restaurant, automatic soft drink filling machines have been installed. For 20-ounce cups, the machine is set to fill up the cups with 19 ounces of soda. Unfortunately, the machine is not perfectly consistent and does not always dispense 19 ounces of soda. Suppose the amount it dispenses produces a normal distribution with a mean of 19 ounces and a standard deviation of 0.6 ounces. It turns out that the 20 ounce cup will actually hold a bit more than 20 ounces. A mathematically inclined worker notices this and starts to record what happens when the machine fills the cups. It turns out that the cups overfill 2% of the time. How much soda will the 20-ounce cup actually hold?

Review Exercises

8) Adult male American bald eagles have a mean wingspan of 79 inches with a standard deviation of 3.5 inches. What percent of these eagles have wingspans longer than 7 feet?



9) Consider the data in the table below where 'x' represents the number of alcoholic drinks consumed and

'y' represents the blood-alcohol concentration.

x	3	1	4	5	3	4
y	0.08	0.02	0.07	0.11	0.05	0.09
x	2	7	8	5	4	8
y	0.06	0.13	0.15	0.07	0.10	0.14

a) Make a labeled and scaled scatterplot for the data set.

b) Determine the correlation coefficient, r, for the scatterplot.

c) Give the least-squares linear regression equation for the scatterplot.

d) Using your answer from part c), predict the blood-alcohol concentration for a person who has had 6 drinks.

e) Using your answer from part c), predict how many drinks someone has had if their blood-alcohol concentration is 0.12.

10) Consider a standard set of 15 pool balls. Pool balls #1-#8 are solid and pool balls #9-#15 are striped.

a) If you randomly select one pool ball, what is the probability that it is both solid and odd?

b) If you randomly select one pool ball, what is the probability that it is either solid or odd?

c) If you randomly select two pool balls without replacement, what is the probability that they are either both solid or both striped?

7.4 Chapter 7 Review

In this chapter we have discussed what a density curve is and specifically focused on a special density curve called the normal distribution. The two critical elements that are necessary for analysis of a density curve are the mean and standard deviation. The mean is the center of the distribution while standard deviation is a measure of spread. We have focused on several key concepts including the 68-95-99.7 rule and z-scores. We then introduced the Normal Distribution Table and the NormalCdf and InvNorm commands to help us be able to move back and forth between probabilities and percentiles and specific values in our distributions.

Chapter 7 Review Exercises

1) Suppose a teacher gives a test in which the scores on the test are normally distributed with a mean of 10 points and a standard deviation of 2 points.

a) Draw a normal curve to represent this situation. Clearly mark the mean and 1, 2, and 3 standard deviations above and below the mean.

b) Using the 68-95-99.7 rule, approximately what percent of students will get a score between 6 and 14?

c) Using the 68-95-99.7 rule, approximately what percent of students will get a score between 8 and 16?

d) Find the percent of students that will get a score between 8 points and 13 points on this test.

e) What percent of students will score at least an 11 points on this test?

f) What percent of students will score between 5 points and 12 points on this test?

g) How many points would a student have to score in order to be at the 90th percentile on this test?

h) What is the z-score associated with a test score of 13 points?

i) How many points did a student score if their z-score was -1.5?

2) Which situation below is most likely to be normally distributed?

i) The heights of all the trees in a forest.

ii) The distances that all the kids at Blaine High School can hit a golf ball.

iii) The number of siblings that each student at Anoka High School has.

iv) The length of time that 6th grade boys at Roosevelt Middle School can hold their breath.

3) The weights of adult male African elephants are normally distributed with a mean weight of 11,000 pounds and standard deviation of 900 pounds.



a) Between what two weights do the middle 50% of all adult male African elephants weigh?

b) Suppose one of these elephants weighs 13,400 pounds. At what percentile is this weight?

c) At what weight would we find the 70th percentile of weights for these elephants?

4) Suppose that IQ test scores are normally distributed with a mean of 100 and a standard deviation of 15.

a) What z-score is associated with and IQ score of 125?

b) The intelligence organization MENSA requires that members score in the top 2.5% of all IQ test takers to gain membership in the organization. What IQ score must a person score to qualify for MENSA?

c) What percentage of IQ scores are greater than 125?

d) What percentage of IQ scores are less than 70? Use the 68-95-99.7 rule to approximate your answer.

e) Who did better, a person with an IQ score of 143 or someone who was at the 99th percentile on the IQ test? Justify your answer.

5) In a certain city, the number of pounds of newspaper recycled each month by a household produces a normal distribution with a mean of 8.5 pounds and a standard deviation of 2.7 pounds.

a) Draw a sketch for this normal distribution and shade in the region that represents the households that recycle between 6 and 12 pounds of newspaper each month.

b) What percent of households recycle between 6 and 12 pounds of newspaper each month?

c) A local newspaper wants to do a story on newspaper recycling in the city. They decide that they would like to base their story on a typical household. After some thought, they decide that 'typical' means that they are in the middle 60% of all households in terms of newspaper recycling. Between what two weights are the 'typical' households?

6) Snowfall each winter in the Twin Cities is normally distributed with a mean of 56 inches and a standard deviation of 11 inches.



a) In what percentage of years does the Twin Cities get less than 3 feet of snow?

b) In what percentage of years does the Twin Cities get more than 6 feet of snow?

c) The winter of 2010-2011 was the fifth snowiest on record for the Twin Cities with a total snowfall of 85 inches. What percentage of years will have snowfalls of more than 85 inches?

d) A winter is considered to be dry if it in the lowest 10% of snowfall totals. What is the maximum amount of snow the Twin Cities could receive to still be called a dry winter?

7) You just got your history test back and found out you scored 37 points. The scores were normally distributed with a mean of 31 points and a standard deviation of 4 points. When you tell your parents how you did, your little brother pipes in that he got a 56 on his math test which was normally distributed with a mean of 40 points and a standard deviation of 11 points. How could you use z-scores to explain to your parents that your score was more impressive than your little brother's score?

8) In 1941, Ted Williams batted 0.406 for the baseball season. He is the last player to hit over 0.400 for an entire major league baseball season. In 2009, Joe Mauer hit 0.365 for the baseball season. In 1941, the batting averages were normally distributed with a mean of 0.260 and a standard deviation of 0.041. In 2009, the batting averages were normally distributed with a mean of 0.262 and a standard deviation of 0.035. Decide which player had a better season compared to the rest of the league during their respective year by comparing z-scores.



9) Suppose that medals will be given out to any student at Andover High School that scores at least 200 points on an aptitude test. The mean score on the aptitude test is 150 points with a standard deviation of 22 points. How many medals should be ordered if there are 456 students who sign up for the test?

Image References

Density Curve www.madscientist.blogspot.com

Skewed Distributions http://en.wikipedia.org/wiki/Skewness

68-95-99.7 Normal Curve www.rahulgladwin.com

Earthworms http://www.flowers.vg

Pet Store Window www.teddyhilton.com

Traffic Jam www.rnw.nl

Leafcutter Ant www.orkin.com/ants

Pair of Queens http://www.123rf.com

American Diabetes Association http://americandiabetesassn.wordpress.com

Track Race http://www.tierraunica.com

Eagle http://www.esa.org

Elephant http://animals.nationalgeographic.com

Blizzard http://www.csc.cs.colorado.edu

Joe Mauer http://www.mauersquickswing.com

http://facstaff.unca.edu/dohse/Stat225/Images/Table-z.JPG

Chapter 8

Appendices

8.1 Appendix A - Tables

Appendix A, Part 1 - Random Digit Table

Line 101 19223 95034 05756 28713 96409 12531 42544 82853
Line 102 73676 47150 99400 01927 27754 42648 82425 36290
Line 103 45467 71709 77558 00095 32863 29485 82226 90056
Line 104 52711 38889 93074 60227 40011 85848 48767 52573
Line 105 95592 94007 69971 91481 60779 53791 17297 59335
Line 106 68417 35013 15529 72765 85089 57067 50211 47487
Line 107 82739 57890 20807 47511 81676 55300 94383 14893
Line 108 60940 72024 17868 24943 61790 90656 87964 18883
Line 109 36009 19365 15412 39638 85453 46816 83485 41979
Line 110 38448 48789 18338 24697 39364 42006 76688 08708
Line 111 81486 69487 60513 09297 00412 71238 27649 39950
Line 112 59636 88804 04634 71197 19352 73089 84898 45785
Line 113 62568 70206 40325 03699 71080 22553 11486 11776
Line 114 45149 32992 75730 66280 03819 56202 02938 70915
Line 115 61041 77684 94322 24709 73698 14526 31893 32592
Line 116 14459 26056 31424 80371 65103 62253 50490 61181
Line 117 38167 98532 62183 70632 23417 26185 41448 75532
Line 118 73190 32533 04470 29669 84407 90785 65956 86382
Line 119 95857 07118 87664 92099 58806 66979 98624 84826
Line 120 35476 55972 39421 65850 04266 35435 43742 11937
Line 121 71487 09984 29077 14863 61683 47052 62224 51025
Line 122 13873 81598 95052 90908 73592 75186 87136 95761

Line 123 54580 81507 27102 56027 55892 33063 41842 81868

Line 124 71035 09001 43367 49497 72719 96758 27611 91596

Line 125 96746 12149 37823 71868 18442 35119 62103 39244

Line 126 96927 19931 36089 74192 77567 88741 48409 41903

Line 127 43909 99477 25330 64359 40085 16925 85117 36071

Line 128 15689 14227 06565 14374 13352 49367 81982 87209

Line 129 36759 58984 68288 22913 18638 54303 00795 08727

Line 130 69051 64817 87174 09517 84534 06489 87201 97245

Line 131 05007 16632 81194 14873 04197 85576 45195 96565

Line 132 68732 55259 84292 08796 43165 93739 31685 97150

Line 133 45740 41807 65561 33302 07051 93623 18132 09547

Line 134 27816 78416 18329 21337 35213 37741 04312 68508

Line 135 66925 55658 39100 78458 11206 19876 87151 31260

Line 136 08421 44753 77377 28744 75592 08563 79140 92454

Line 137 53645 66812 61421 47836 12609 15373 98481 14592

Line 138 66831 68908 40772 21558 47781 33586 79177 06928

Line 139 55588 99404 70708 41098 43563 56934 48394 51719

Line 140 12975 13258 13048 45144 72321 81940 00360 02428

Line 141 96767 35964 23822 96012 94591 65194 50842 53372

Line 142 72829 50232 97892 63408 77919 44575 24870 04178

Line 143 88565 42628 17797 49376 61762 16953 88604 12724

Line 144 62964 88145 83083 69453 46109 59505 69680 00900

Line 145 19687 12633 57857 95806 09931 02150 43163 58636

Line 146 37609 59057 66967 83401 60705 02384 90597 93600

Line 147 54973 86278 88737 74351 47500 84552 19909 67181

Line 148 00694 05977 19664 65441 20903 62371 22725 53340

Line 149 71546 05233 53946 68743 72460 27601 45403 88692

Line 150 07511 88915 41267 16853 84569 79367 32337 03316

Appendix A, Part 2 - The Normal Distribution Table

For z-scores with z less than or equal to zero

Table 8.1:

z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.0
-3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357

Table 8.1: (continued)

-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
-0.2	0.3829	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
-0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
-0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

For z-scores with z greater than or equal to 0

Table 8.2:

z	**0.0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0..6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133

Table 8.2: (continued)

0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Appendix A, Part 3 - A standard deck of 52 cards

Clubs	Spades	Hearts	Diamonds
A♣	A♠	A♥	A♦
2♣	2♠	2♥	2♦
3♣	3♠	3♥	3♦
4♣	4♠	4♥	4♦
5♣	5♠	5♥	5♦
6♣	6♠	6♥	6♦
7♣	7♠	7♥	7♦
8♣	8♠	8♥	8♦
9♣	9♠	9♥	9♦
10♣	10♠	10♥	Jack♦
Jack♣	Jack♠	Jack♥	Queen♦
Queen ♣	Queen♠	Queen♥	King♦
King♣	King♠	King♥	King♦

Figure 8.1

Appendix A, Part 4 - Results for the total of two 6-sided dice

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Figure 8.2

8.2 Appendix B - Glossary and Index

95% confidence statement - Page , Section

"We are 95% confident that the true proportion of _____(parameter of interest)_____ will be between _____(low value of conf. int.)_____ and _____(high value of conf. int.)_____."

Back to Back Stem Plots - Page , Section

A stem plot in which two sets of numerical data share the stems in the middle, with one set has its leaves going to the right and the other set has its leaves going to the left.

Bar Graph - Page , Section

A graph in which each bar shows how frequently a given category occurs. The bars can go either horizontally or vertically. Bars should be of consistent width and need to be equally spaced apart. The categories may be placed in any order along the axis.

Bias - Page , Section

A measurement that is repeatedly either too high or too low.

Bin Width

See Class Size

Bi-variate Data - Page , Section

Numerical data that measures two variables.

Blinded Study - Page , Section

A study in which the subject does not know exactly what treatment they are getting.

Block Design - Page , Section

A study in which subjects are divided into distinct categories with certain characteristics (for example, males and females) before being randomly assigned treatments.

Box Plot (Box and Whisker Plot) - Page , Section

A display in which a numerical data set is divided into quarters. The 'box' marks the middle 50% of the data and the 'whiskers' mark the upper 25% and lower 25% of the data.

Categorical Variable - Page , Section

Variables that can be put into categories, like favorite color, type of car you own, your sports jersey number, etc...

Census - Page , Section

A special type of study in which data is gathered from every single member of the population.

Center - Page , Section

Typically, it is the mean, median, or the mode of a data set. In a normal distribution curve the mean, median, and mode all mark the center.

Chance Behavior - Page , Section

Events whose outcomes are not predictable in the short term, but have long term predictability.

Class Size (Bin Width) - Page , Section

A consistent width that all bars on a histogram have. A quick estimation of a reasonable class size is to roughly divide the range by a value from about 7 to 10.

Coincidence - Page , Section

A relationship between two variables that simply occurs by chance.

Combination - Page , Section

An arrangement of a set of object in which the order does not matter. ${}_nC_r = \frac{n!}{r!(n-r)!}$

Common Response - Page , Section

A situation in which two variables have similar behaviors but are actually both responding to an additional lurking variable.

Complement of an Event - Page , Section

The probability of an event, 'A', NOT occurring. It can be thought of the opposite of an event and can be notated as A^c or ˜A.

Compound Event - Page , Section

An event with two or more steps such as drawing a card and then rolling a die.

Conditional Probability - Page , Section

The probability of a particular outcome happening assuming a certain prerequisite condition has already been met. A clue that a conditional probability is being considered is the word 'given' or the vertical bar symbol, |.

Confidence Interval - Page , Section

The range of answers included within the margin of error. Typically, we use a 95% confidence interval meaning it is very likely (95% chance) that the parameter lies within this range.

Confounding - Page , Section

Occurs when two variables are related, but it is not a clear cause/effect relationship because there are other variables that are carrying influence in the situation.

Context - Page , Section

The specific realities of the situation we are considering. We often consider the labels and units when defining the context.

Contingency Table

See 2-Way Table

Control Group - Page , Section

A group in an experiment that does not receive the actual treatment, but rather receives a placebo, a known treatment, or no treatment at all.

Convenience Sample - Page , Section

A biased sampling method in which data is only gathered from those individuals who are easy to ask or are conveniently located.

Correlation (r) - Page , Section

A statistic that is used to measure the strength and direction of a linear correlation whose values range from -1 to 1. The sign of the correlation (+/-) matches the sign of the slope of the regression equation.

Data - Page , Section

A collection of facts, measurements, or observations about a set of individuals.

Density Curve - Page , Section

A curve that gives a rough description of a distribution. The curve is smooth and always has an area equal to 1 whole or 100%.

Dependent Events - Page , Section

A situation in which one event changes the probability of another event.

Direct Cause and Effect - Page , Section

A situation in which one variable causes a specific effect to occur with no lurking variables.

Direction - Page , Section

One of three general results reported for a linear regression. It will be reported as either be positive, negative, or 0.

Disjoint

See Mutually Exclusive Events

Dot Plot - Page , Section

A simple display that places a dot above the axis for each value. There is a dot for each value, so values that occur more than once will be shown by stacked dots.

Double Blind - Page , Section

A study in which neither the experimenter nor the subject knows which treatment is being given.

Empirical Rule (68-95-99.7 Rule) - Page , Section

A rule that states that in a normal distribution, 68% of the data is located within one standard deviation from the mean, 95% of the data is located within two standard deviations from the mean, and 99.7% of the data is located within three standard deviations from the mean.

Event - Page , Section

Any action from which a result will be recorded or measured.

Expected Value - Page , Section

The average result over the long run for an event if repeated a large number of times.

Experiment - Page , Section

A study in which the researchers impose a treatment on the subjects.

Explanatory Variable - Page , Section

The x-axis variable. It can often be viewed as the 'cause' variable or the independent variable.

Factorial - Page , Section

A number followed by an exclamation point indicated repeated multiplication down to 1. For example, $4! = 4 \times 3 \times 2 \times 1$.

Fair Game - Page , Section

A game in which neither the player nor the house has an advantage. An average player over the long run will neither gain nor lose money. In other words, the expected value of the game is the same as the cost to play the game.

Five-Number Summary - Page , Section

A description of data that includes the minimum, first quartile, median, third quartile, and maximum numbers which can be used to create a box plot.

Form - Page , Section

A general description of the pattern in a scatterplot. Typical descriptions include linear, curved, or random (no specific form).

Frequency Table - Page , Section

A table that shows the number of occurrences in each category.

Fundamental Counting Principle - Page , Section

A rule that states to find the number of outcomes for a given situation, simply multiply the number of outcomes for each individual event.

Histogram - Page , Section

A special bar graph for a numerical data set. In a histogram, each bar has the same width with no space between them where bars track the frequency of results in its given range.

Independent Events - Page , Section

Two events in which the outcome of one event does not change the probabilities for the outcome for the other event.

Individual - Page , Section

The subject being studied. This can be a person, an animal or an object.

Inter-Quartile Range (IQR) - Page , Section

The distance between the lower and upper quartiles. IQR = Q_3 - Q_1

Instrument of Measurement - Page , Section

Tool used to make measurements. Typical instruments are tools like rulers, scales, thermometers, or speedometers.

Intersection of Sets - Page , Section

In a Venn Diagram, it includes the results that are members of more than one group simultaneously. We use the symbol, $\cap$, to indicate the intersection and think of the intersection of those parts of the diagram that include both A and B.

Law of Large Numbers - Page , Section

A rule that states that we will eventually get closer to the theoretical probability as we greatly increase the number of times an event is repeated.

Line Graph

See Time Plot

Lurking Variable - Page , Section

An additional variable that was not taken into account in a particular situation.

Margin of Error - Page , Section

A range of results, often spanning from 2 standard deviations below to 2 standard deviations above the mean in which we are 95% confident that the true parameter is located. The quick method for an approximation of the margin of error for a 95% confidence interval is $M.O.E = \frac{1}{\sqrt{n}}$.

Mean (Average) - Page , Section

The sum of all the numbers divided by the number of values in the data set. It is also located at the center of a normal distribution and is a good measure of center for symmetric data sets.

Median - Page , Section

The data result in the middle of a data list that has been organized smallest to largest. If there are two middle data values, then the median is located halfway between those two values. In a visual distribution, it marks the 50/50 area point on the graph. Use for skewed data sets.

Mode - Page , Section

The result that appears most frequently in a data set. It also occurs at the highest point of a density curve.

Multistage Random Sample - Page , Section

A sampling technique that uses randomly selected sub-groups of a population before random selection of individuals occurs.

Mutually Exclusive Events (Disjoint) - Page , Section

Events that cannot occur at the same time.

Negative Linear Association - Page , Section

A situation such that as one numerical variable increases, another numerical variable decreases.

Non-Response - Page , Section

A non-sampling error in which subjects do not participate or do not answer questions in a survey.

Normal Distribution Curve - Page , Section

A bell-shaped curve that describes a symmetrical data set such that the most frequent results occur near the mean and results become less frequent as you move further from the mean.

Numerical Variable - Page , Section

A variable that can be assigned a numerical value, such as a height, a distance, a temperatures, etc...

Observational Study - Page , Section

A study in which researchers do not impose a treatment on the subjects. Data is collected by watching the subjects or from information already available. (Observe but do not disturb)

Outcome - Page , Section

A possible result of an event.

Outlier - Page , Section

A value that is unusual when compared to the rest of a data set. High outliers will be greater than Q_3+1.5IQR. Low outliers will be below Q_1-1.5IQR.

Parallel box plots - Page , Section

Multiple box plots graphed on the same axes to compare multiple data sets.

Parameter - Page , Section

A value that describes a truth about a population. Sometimes, the value is unknown so a parameter is often given as a description of truth.

Permutation - Page , Section

A specific order or arrangement of a set of objects or items. In a permutation, the order in which the items are selected matters.

Pictograph - Page , Section

A bar graph that uses pictures instead of bars. These graphs can be misleading because pictures measure height and width, where bar graphs measure only height. To be effective, all the pictures used must be the same size.

Pie chart - Page , Section

A graph which shows each category as a part of the whole in a circle graph. Pie charts can be used if exactly 100% of the results for a particular situation are known.

Placebo - Page , Section

A fake treatment that is similar in appearance to the real treatment.

Placebo Effect - Page , Section

The placebo effect occurs when a subject starts to experience changes simply because they believe they are receiving a treatment.

Population - Page , Section

The entire group of individuals we are interested in.

Positive Linear Association - Page , Section

A situation such that as one numerical variable increases, the other numerical variable also increases.

Prime Number - Page , Section

A number that is divisible only by 1 and itself. Remember, 1 is not a prime number!

Probability - Page , Section

The likelihood of a particular outcome occurring.

Probability Model - Page , Section

A table that lists all outcomes of an event and their respective probabilities. The sum of all the probabilities in a probability model must equal 1.

Processing Errors - Page , Section

An error commonly made due to issues like poor calculations or inaccurate recording of results.

Prospective Studies - Page , Section

A study which follows up with study subjects in the future in an effort to see if there were any long-term effects.

Quartile 1 - Page , Section

The median of all the values to the left of the median. Do not include the median itself.

Quartile 3 - Page , Section

The median of all the values to the right of the median. Do not include the median itself.

Random Event - Page , Section

An event for which we can not be certain of the outcome.

Table of Random Digits - Page , Section

A long list of randomly chosen digits from 0 to 9, usually generated by computer software or calculators. A table of random digits can be found in Appendix A, Part 1.

Random Sampling Error - Page , Section

Even though a sample is randomly selected, it is entirely possible that a particular result within the population will be over-represented. Larger sample sizes reduce random sampling error. The margin of error is stated with most studies to account for random sampling error.

Range - Page , Section

A basic description of how spread out a data set is. It is calculated by subtracting the smallest number in a data set from the largest number in the data set.

Reliability - Page , Section

How consistently a particular measurement technique gives the same, or nearly the same measurement.

Response Bias - Page , Section

Occurs when an individual responds to a survey with an incorrect or untruthful answer. This type of bias can frequently happen when questions are potentially sensitive or embarrassing.

Response Variable - Page , Section

The y-axis variable. It can often be thought of as the 'effect' variable or dependent variable.

Retrospective Study - Page , Section

A study in which information about a subject's past is used in the study.

Sample - Page , Section

A representative subset of a population.

Sample Space - Page , Section

A list of all the possible outcomes that may occur.

Sample Survey - Page , Section

A survey that uses a subset of the population in order to try to make predictions about the entire population.

Sampling Frame - Page , Section

A list of all members of a population.

Scatterplot - Page , Section

Graphs that represent a relationship between two numerical variables where each data point is shown as a coordinate point on a scaled grid.

SCOFD

This is used for the description of a scatterplot and stands for Strength, Context, Outliers, Form, and Direction.

Simple Random Sample (SRS) - Page , Section

A sample where all possible groups of a particular size are equally possible. It can be thought of as putting all members of the population in a hat and randomly drawing until the desired sample size is reached.

Simulation - Page , Section

A model of a real situation that can be used to make predictions about what might really happen. Often, tables of random digits are used to carry out simulations.

Skewed Distribution - Page , Section

A distribution in which the majority of the data is concentrated on one end of the distribution. Visually, there is a 'tail' on the side with less data and this is the direction of the skew.

SOCCS - Page , Section

A way to remember the key information to discuss for a distribution: Shape, Outliers, Center, Context, and Spread.

Spread - Page , Section

A way to measure variability of a data set. Common measures of spread are the range, standard deviation, and IQR.

Standard deviation - Page , Section

A measure of spread relative to the mean of a data set. Use this measurement for any data set which is approximately normally distributed.

Statistic - Page , Section

A number that describes results from sample. This number is often used to make an approximation of the parameter.

Stem plot - Page , Section

A method of organizing data that sorts the data in a visual fashion. The stem is made up of all the leading digits of a piece of data and the leaf is the final digit.

Stratified Random Sample - Page , Section

A sample in which the population is divided into distinct groups called strata before a random sample is chosen from each strata.

Strength - Page , Section

One of three measurements reported for a best-fit line that describes how closely the data matches a perfect line.

Subjects - Page , Section

The individuals that are being studied in an experiment.

Symmetrical Distribution - Page , Section

A distribution in which the left side of the distribution looks like the mirror image of the right side of the distribution.

Systematic random sample - Page , Section

A sampling method in which the first selection is made randomly and then a 'system' is used to make the remaining selections.

Theoretical Model - Page , Section

A model that gives a picture of exactly the frequencies of what should happen in a situation involving probability.

Theoretical probability - Page , Section

A mathematical calculation of the likelihood that an event will occur.

Time Plot (Line Graph) - Page , Section

A graph that shows how a variable changes over time.

Tree Diagram - Page , Section

A visual representation of a series of events where each successive event branches off from the previous event.

Two-Way Table (Contingency Table) - Page , Section

A table which tracks two characteristics from a set of individuals. For example, we might track gender and grade of all the students in your high school.

Undercoverage - Page , Section

A sampling error in which an entire group or groups of subjects are left out or underrepresented in a study.

Union of Sets - Page , Section

A union includes all results that are in either one category, another category, or both categories in a Venn diagram. We use the symbol $\cup$ and can think of a union as either A or B (or both).

Validity - Page , Section

A measurement technique is valid if it is an appropriate way to collect data.

Variables - Page , Section

Characteristics about the individuals that the researchers might be interested in.

Venn Diagrams - Page , Section

Diagrams that represent outcomes using intersecting circles.

Voluntary Response Sample - Page , Section

A biased sampling method in which participants get to choose whether or not to participate in the survey. The bias occurs because those who are most passionate about an issue will be more likely to respond.

Wording of a Question - Page , Section

The wording of a question can be used to manipulate subjects to make them more likely to respond a certain way in a survey causing bias.

Z-Score - Page , Section

A measure of the number of standard deviations a particular data point is away from the mean in a normal distribution. If a z-score is positive, the value is larger than the mean and if it is negative, it is less than the mean.

8.3 Appendix C - Calculator Help

This appendix is not meant to be a full guide for calculators common to students who take this course. Rather, it is intended to highlight some of the key procedures used on a variety of different calculators. The steps are arranged by topic as opposed to being arranged by calculator. One online source that is helpful for those of you with graphing calculator issues can be found on the Prentice Hall website at http://www.prenhall.com/divisions/esm/app/calc_v2/ .

Topic 1 - Combinations and Permutations

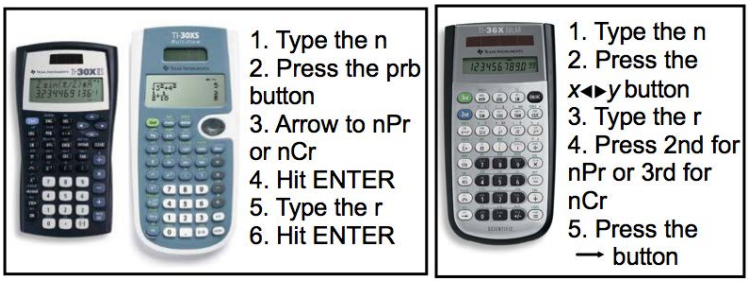

1. Type the n
2. Press the prb button
3. Arrow to nPr or nCr
4. Hit ENTER
5. Type the r
6. Hit ENTER
1. Type the n
2. Press the x◄►y button
3. Type the r
4. Press 2nd for nPr or 3rd for nCr
5. Press the → button

Figure 8.3

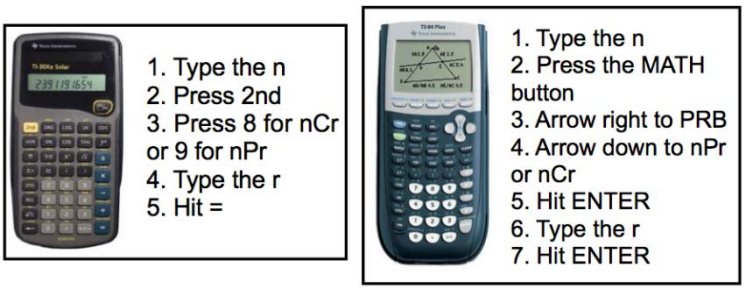

1. Type the n
2. Press 2nd
3. Press 8 for nCr or 9 for nPr
4. Type the r
5. Hit =
1. Type the n
2. Press the MATH button
3. Arrow right to PRB
4. Arrow down to nPr or nCr
5. Hit ENTER
6. Type the r
7. Hit ENTER

Figure 8.4

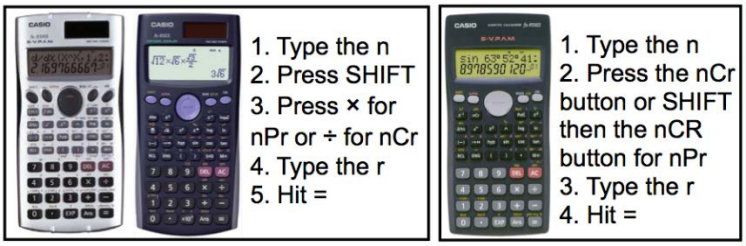
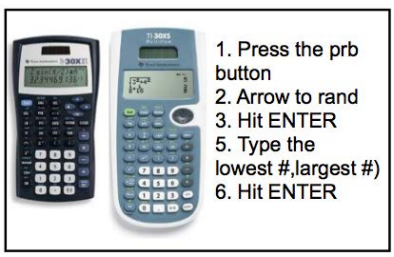

1. Type the n
2. Press SHIFT
3. Press × for nPr or ÷ for nCr
4. Type the r
5. Hit =
1. Type the n
2. Press the nCr button or SHIFT then the nCR button for nPr
3. Type the r
4. Hit =

Figure 8.5

Topic 2 - Random Number Generators



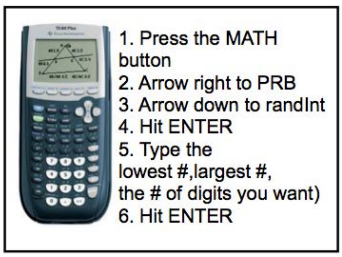

1. Press the MATH button
2. Arrow right to PRB
3. Arrow down to randInt
4. Hit ENTER
5. Type the lowest #,largest #, the # of digits you want)
6. Hit ENTER

Figure 8.6

Topic 3 - Means and Standard Deviations

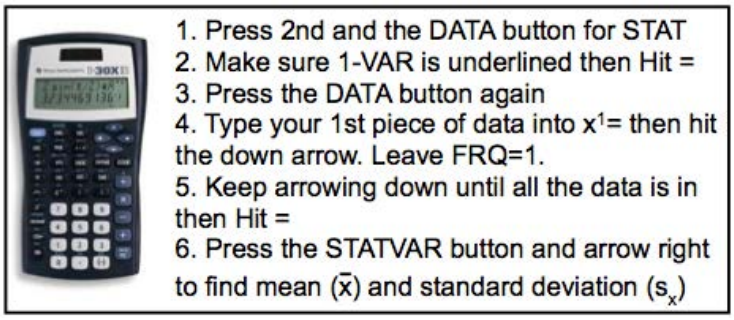

1. Press 2nd and the DATA button for STAT
2. Make sure 1-VAR is underlined then Hit =
3. Press the DATA button again
4. Type your 1st piece of data into x1= then hit the down arrow. Leave FRQ=1.
5. Keep arrowing down until all the data is in then Hit =
6. Press the STATVAR button and arrow right to find mean ($\bar{x}$) and standard deviation (s_x)

Figure 8.7

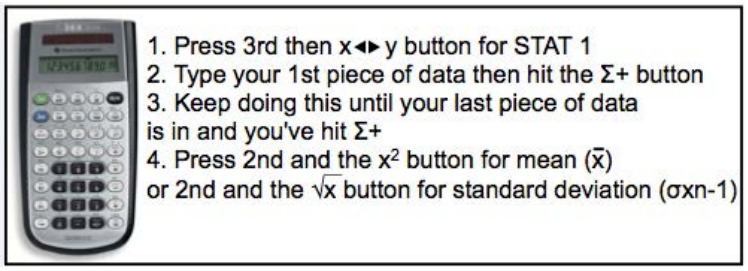

1. Press 3rd then x◂▸y button for STAT 1
2. Type your 1st piece of data then hit the Σ+ button
3. Keep doing this until your last piece of data is in and you've hit Σ+
4. Press 2nd and the x^2 button for mean ($\bar{x}$) or 2nd and the $\sqrt{x}$ button for standard deviation (σxn-1)

Figure 8.8

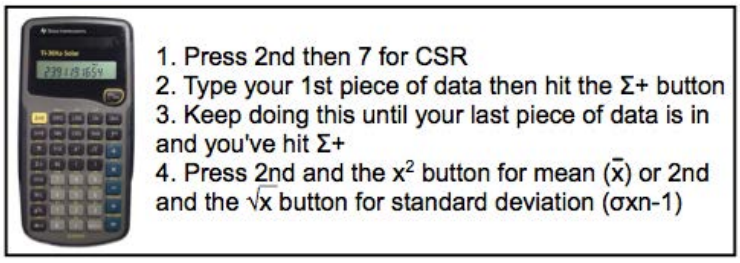
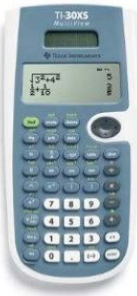

1. Press 2nd then 7 for CSR
2. Type your 1st piece of data then hit the Σ+ button
3. Keep doing this until your last piece of data is in and you've hit Σ+
4. Press 2nd and the x^2 button for mean ($\bar{x}$) or 2nd and the $\sqrt{x}$ button for standard deviation (σxn-1)

Figure 8.9



1. Press the DATA button
2. Type your data into L_1
3. Press 2nd then the DATA button again for STAT
4. Highlight 1:1-Var Stats then hit ENTER
5. Highlight L_1 then hit ENTER
6. Highlight Frq 1 then hit ENTER
7. Highlight CALC then hit ENTER
8. Arrow down to find mean ($\bar{x}$) and standard deviation (s_x)

Figure 8.10



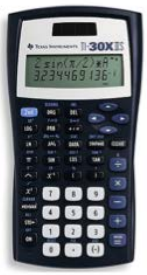
1. Press the STAT button
2. Choose 1: Edit
3. Type your data into L_1
4. Press the STAT button again
5. Arrow to the right for CALC
6. Choose 1: 1-Var Stats
7. Press ENTER again
8. Find mean ($\bar{x}$) and standard deviation (s_x)

Figure 8.11



1. Press the MODE button
2. Choose STAT
3. Press 1 for 1-VAR
4. Type your data in the x list
5. When all the data is in press the AC button
6. Press SHIFT
7. Press 1 for STAT
8. Press 5 for Var
9. Press 2 for mean ($\bar{x}$) or 4 for standard deviation (xσn-1)
10. Hit =

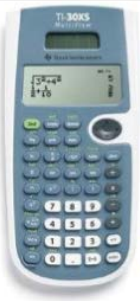
Figure 8.12



1. Press 2nd then the DATA button for STAT
2. Make sure 2-VAR is underlined then Hit =
3. Press the DATA button
4. Type your 1st x-value into x^1= then arrow down and type your 1st y-value into y^1=
5. Keep arrowing down until all your data is entered in then Hit =
6. Press the STATVAR button and arrow right to find slope (a) or y-intercept (b) or correlation (r)

*NOTE: a & b mean the opposite here

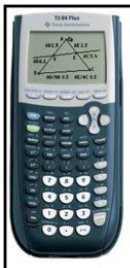
Figure 8.13



1. Press the DATA button
2. Type your x data into L_1 and your y data into L_2
3. Press 2nd then the DATA button again for STAT
4. Highlight 2: 2-Var Stats then hit ENTER
5. Highlight L_1 for x then hit ENTER
6. Highlight L_2 for y then hit ENTER
7. Highlight CALC then hit ENTER
8. Arrow down to find slope (a) or y-intercept (b) or correlation (r)

*NOTE: a & b mean the opposite here

Figure 8.14



1. Press the STAT button
2. Choose 1: Edit
3. Type your x-values into L_1 and your y-values into L_2
4. Press the STAT button again
5. Arrow to the right for CALC
6. Choose 8: LinReg(a+bx)
7. Press ENTER again
8. Find y-intercept (a) or slope (b) or correlation (r)

*NOTE: If r does not appear, then you need to follow these steps once to turn it on.

1. Press 2nd then the 0 for CATALOG
2. Arrow down until DiagnosticOn is highlighted
3. Press ENTER until the screen shows Done

Figure 8.15

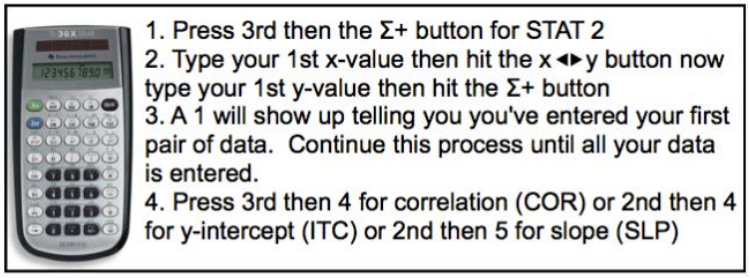

1. Press 3rd then the Σ+ button for STAT 2
2. Type your 1st x-value then hit the x◄►y button now type your 1st y-value then hit the Σ+ button
3. A 1 will show up telling you you've entered your first pair of data. Continue this process until all your data is entered.
4. Press 3rd then 4 for correlation (COR) or 2nd then 4 for y-intercept (ITC) or 2nd then 5 for slope (SLP)

Figure 8.16

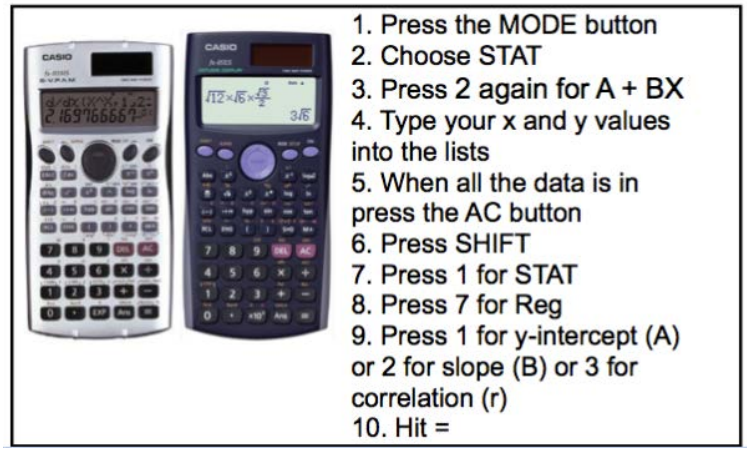

1. Press the MODE button
2. Choose STAT
3. Press 2 again for A + BX
4. Type your x and y values into the lists
5. When all the data is in press the AC button
6. Press SHIFT
7. Press 1 for STAT
8. Press 7 for Reg
9. Press 1 for y-intercept (A) or 2 for slope (B) or 3 for correlation (r)
10. Hit =

Figure 8.17

Topic 5 - Normal Distributions

Your graphing calculator has already been programmed to calculate probabilities for a normal density curve using what is called a cumulative density function or cdf. This is found in the distributions menu above the VARS key.



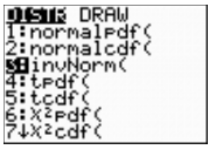
Press **[2nd] [VARS], [2]** to select the normalcdf

To get normalcdf to work properly, you will need to enter the two values that you are looking at followed by the mean and standard deviation. Your screen should look something like the following:

$$\text{normalcdf}\ (a, b, \mu, \sigma)$$

Figure 8.18

Your TI – 83/84 graphing calculators have already been programmed to find values at certain percentiles in a normal curve. This feature is called invNorm and can be found in the Distribution Menu.



Press **[2nd] [VARS], [3]** to select the invNorm

To get invNorm to work properly, you will need to enter the percentile followed by the mean and standard deviation. Your screen should look something like the following (you must enter the percent as a decimal):

$$\text{invNorm}\ (\%, \mu, \sigma)$$

Figure 8.19

Image References

Random Digit Table http://uwsp.edu/math

Normal Distribution Table http://www.regentsprep.org